GESIS Leibniz Institute for the Social Sciences

Meet the Experts @GESIS

Meet the Experts - GESIS online talks

KODAQS Toolbox: Using *SampcompR* for survey representation bias comparisons

Leibniz Association

Unlocking Data Quality with KODAQS: Smart Tools for Better Social Science Data

06.11.2025

Björn Rohr

Presenters/authors

Björn Rohr

- A survey methodologist, who defended his PhD last week and is currently working in a DFG-funded Research Project on Modular Questionnaire Designs and Multiple Imputation

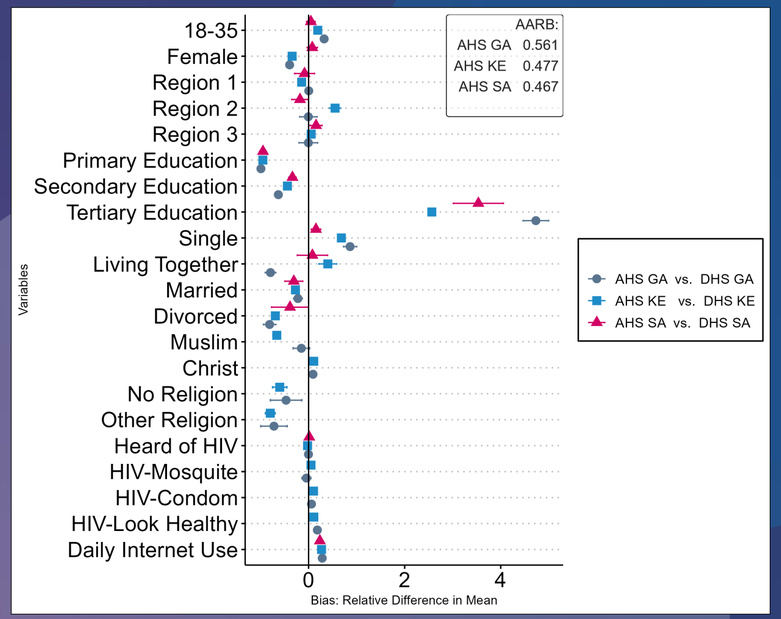
- Research Interests: Representation Bias in Surveys, Survey Recruitment, Nonprobability Surveys, Modular Questionnaire Designs and Multiple Imputations

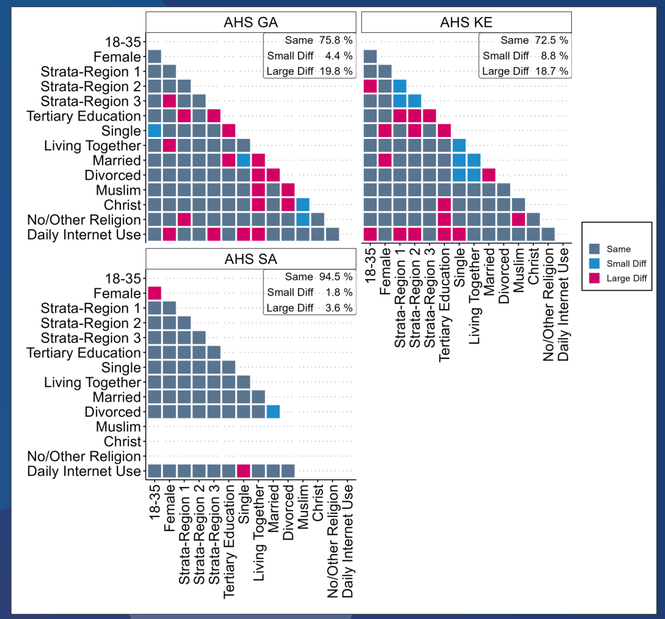
- Contact: bjoern.rohr@gesis.org

GESIS Leibniz Institute for the Social Sciences

What is SampcompR?

Leibniz Association





SampcompR

- SampcompR is an R-Package to estimate differences between samples (i.e., surveys and benchmark (surveys)).
- Differences (e.g., bias) can be estimated on a **univariate**, **bivariate**, and **multivariate** level.
- Estimation can include **weighting**, **bootstrap** & **Bonferroni adjustment**
- Differences can be visualized in a plot or outputted as a table.

Authors:
Björn Rohr (Bjoern.Rohr@gesis.org)
Barbara Felderer (Barbara.Felderer@gesis.org)

GESIS Leibniz Institute for the Social Sciences

Selected Functions of SampcompR

Univariate Functions

- uni_compare()
- plot_uni_compare()
- uni_compare_table()

Bivariate Functions

- biv_compare()
- plot_biv_compare()
- biv_compare_table()

Multivariate Functions

- multi_compare()
- plot_multi_compare()
- multi_compare_table()

GESIS Leibniz Institute for the Social Sciences

Example Data & Benchmark Data

Leibniz Association

GESIS Leibniz Institute for the Social Sciences

Example Surveys

African Health Surveys (AHS)

- **Countries:** Ghana (N = 494), Kenya (N = 2,552), South Africa (N = 275)
- **Target Population:** Every respondent between 18 - 65+ years, living in the chosen countries
- **Method:** Gathered with Facebook ads using 5 different pictures
- **Field Period:** 3 weeks (Wednesday - Friday) between 15.02.2023 - 03.03.2023
- **Budget:** **168 €** per day, country, and targeting strategy
- **Incentives:** Lottery for 2x 5GB mobile data per country.

GESIS Leibniz Institute for the Social Sciences

Benchmark Survey

Demographics & Health Survey (DHS)

Cross-National probability survey in **Ghana** (N = 7,863), **Kenya** (N = 24,003), and **South Africa** (N = 6,708), among other countries

Target Population: Males 18 – 54 years old and females 18 – 49 years old

Field Periods:

- **Ghana:** September 2022 – December 2022
- **Kenya:** May 2022- October 2022
- **South Africa:** June 2016 - November 2016

Sampling Strategy: Two-stage cluster sample (design weights were used in all analyses)

GESIS Leibniz Institute for the Social Sciences

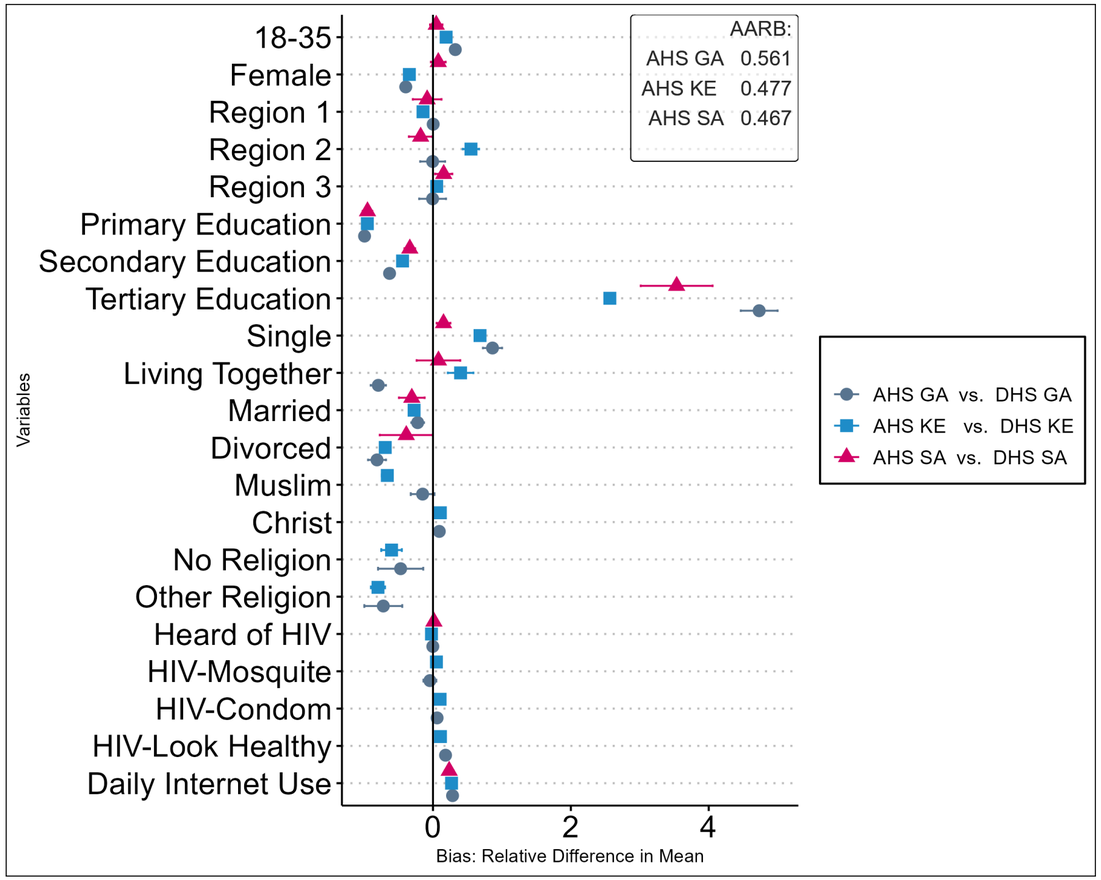
Example Univariate Comparison

Leibniz Association

Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```

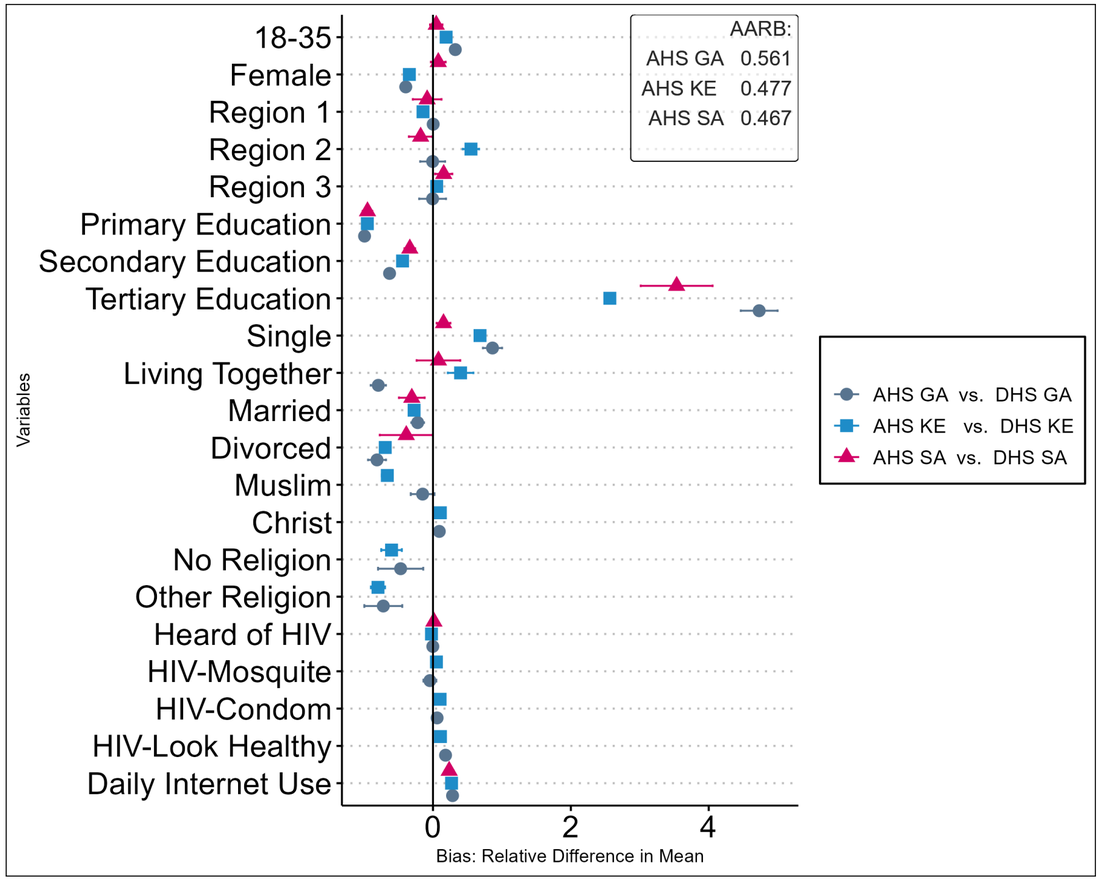


Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```

Three **surveys** are compared against three **benchmark surveys**



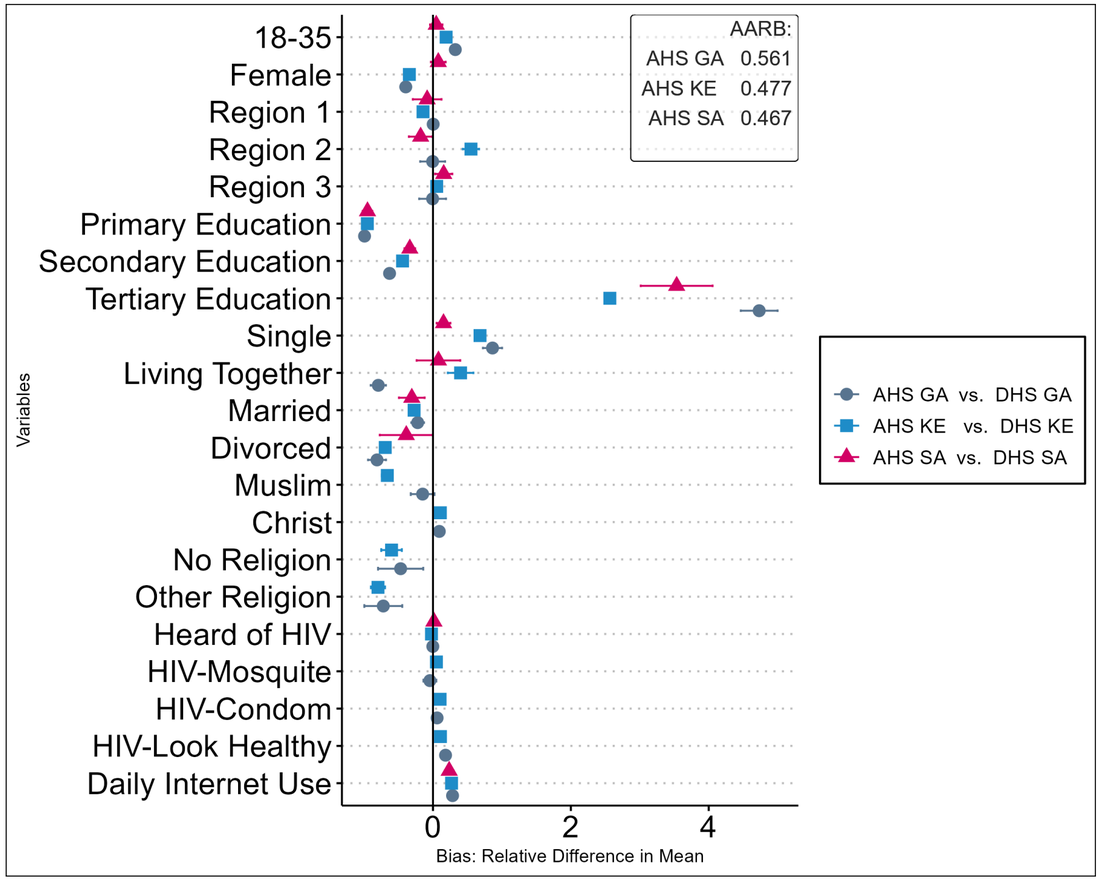
Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```

Bootstrap confidence intervals are calculated with 2000 bootstraps

Alternative: No bootstrap confidence intervals (nboots =0)

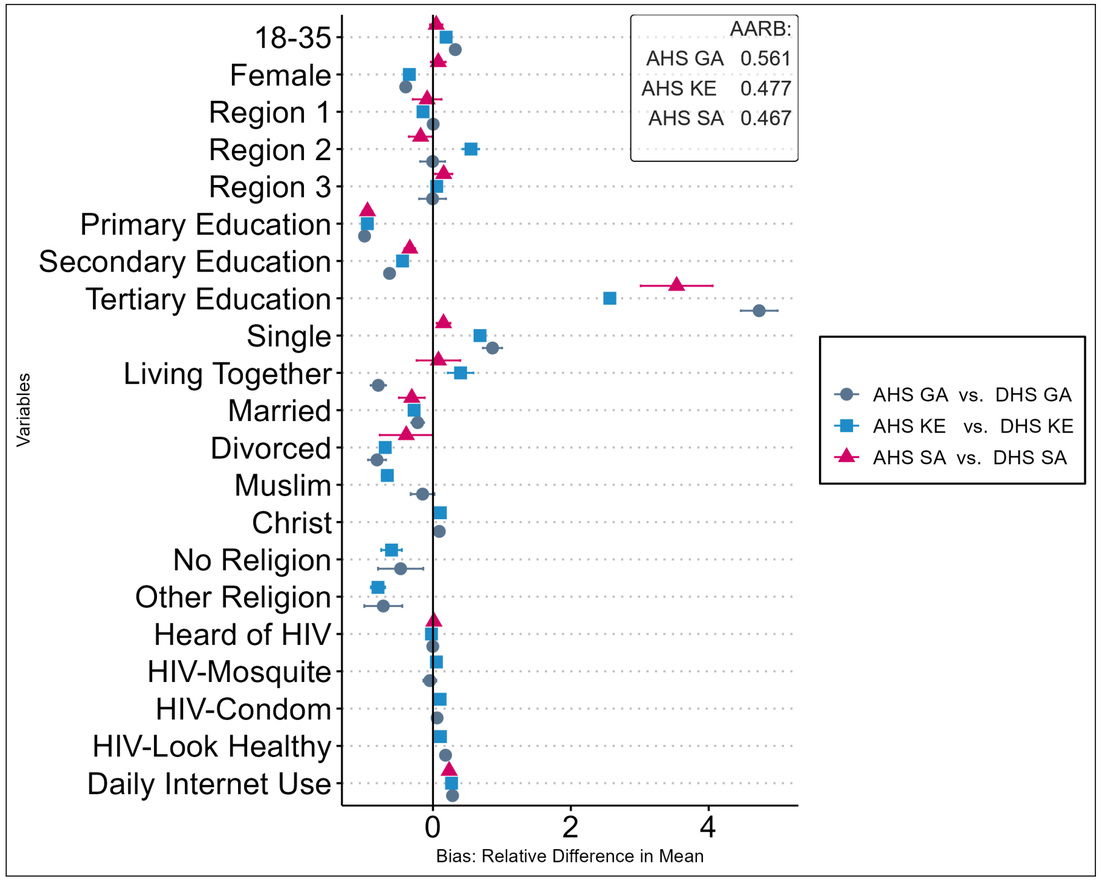


Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),

variables<-c("young", "female", "strata_reg1", "strata_reg2", "strata_reg3",
             "primary_educ2", "secondary_educ","tertiary_educ", "single",
             "living_together", "married", "divorced","muslim", "christ",
             "no_rel", "other_rel","hiv_heard","hiv_mosquito_corr",
             "hiv_condom_corr","hiv_healthy_corr","internet_daily")

                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```

funct = rel_mean → Relative bias in mean is calculated
Alternatives: Absolute Relative bias in mean, (absolute) bias in mean



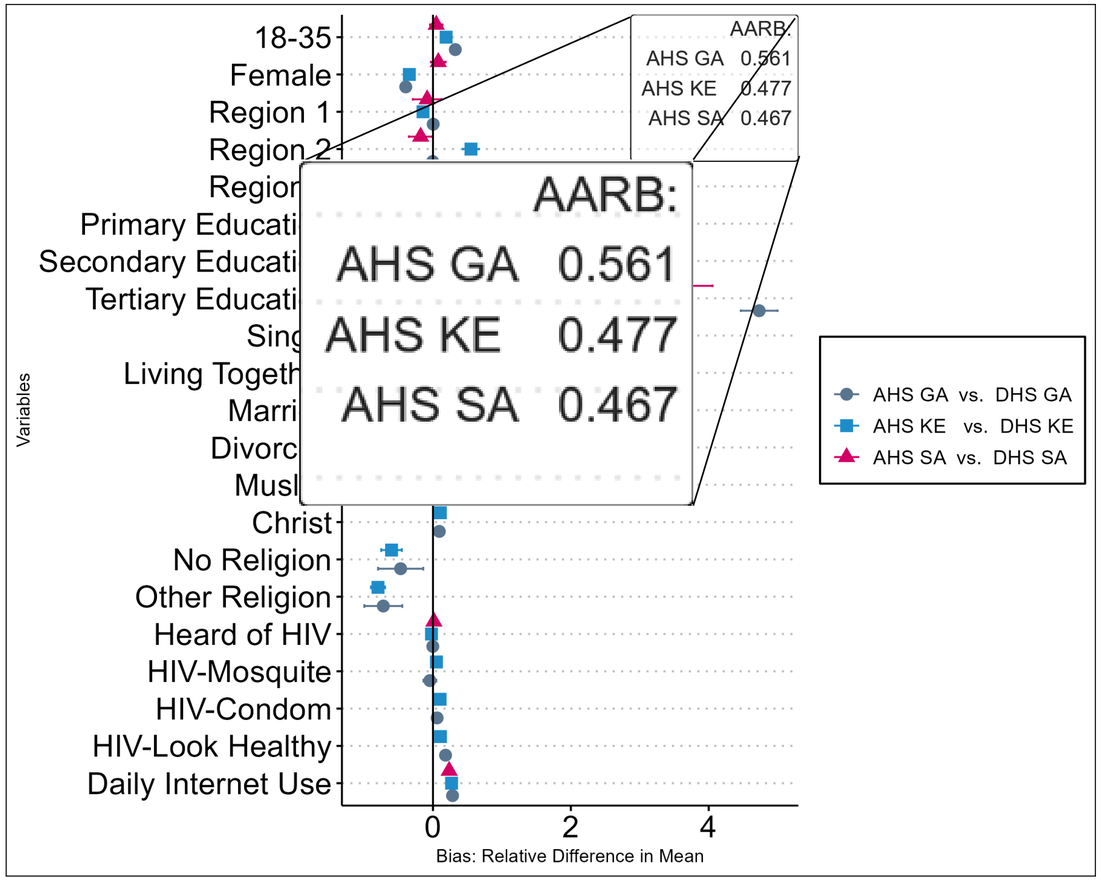
Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```

To summarize the bias in each survey, Average Absolute Relative Bias (AARB) is calculated
Alternatives: RMSE, MSE

The confidence intervals are Bonferroni corrected.
Alternatives: No correction

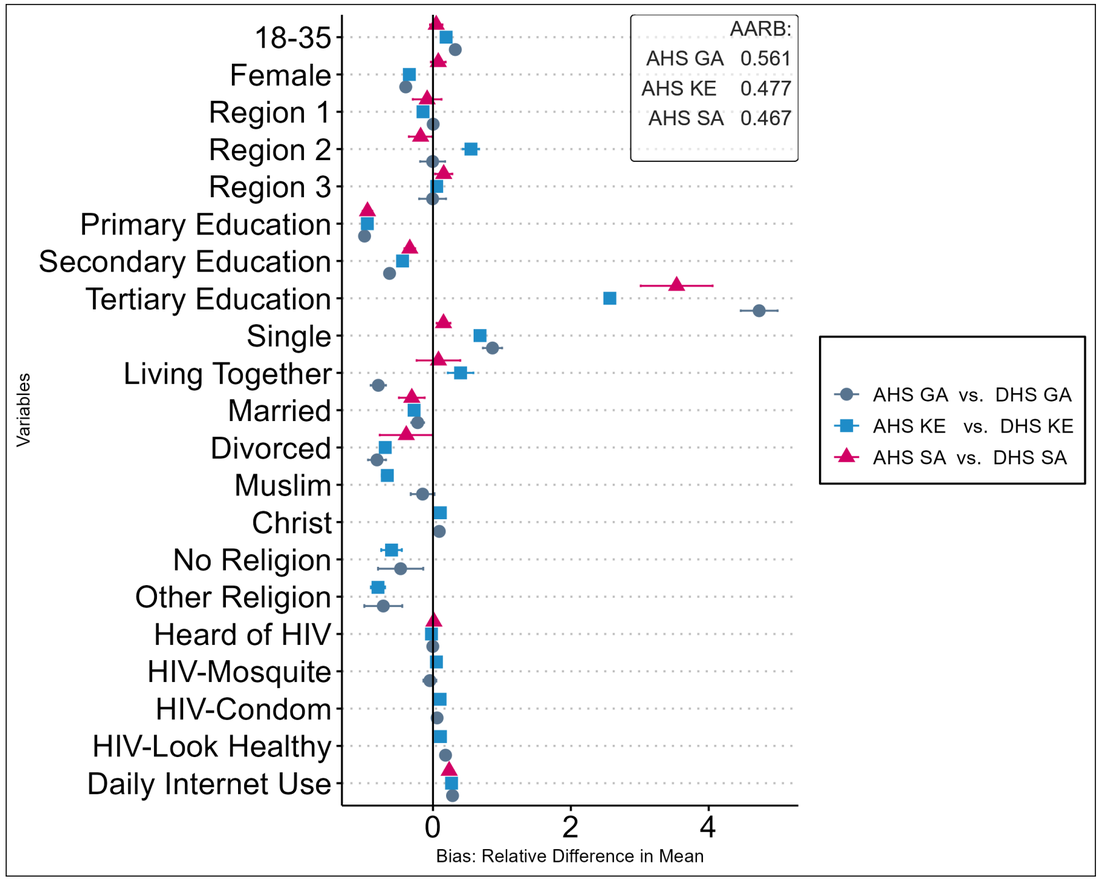


Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```

The Surveys and benchmarks are relabled



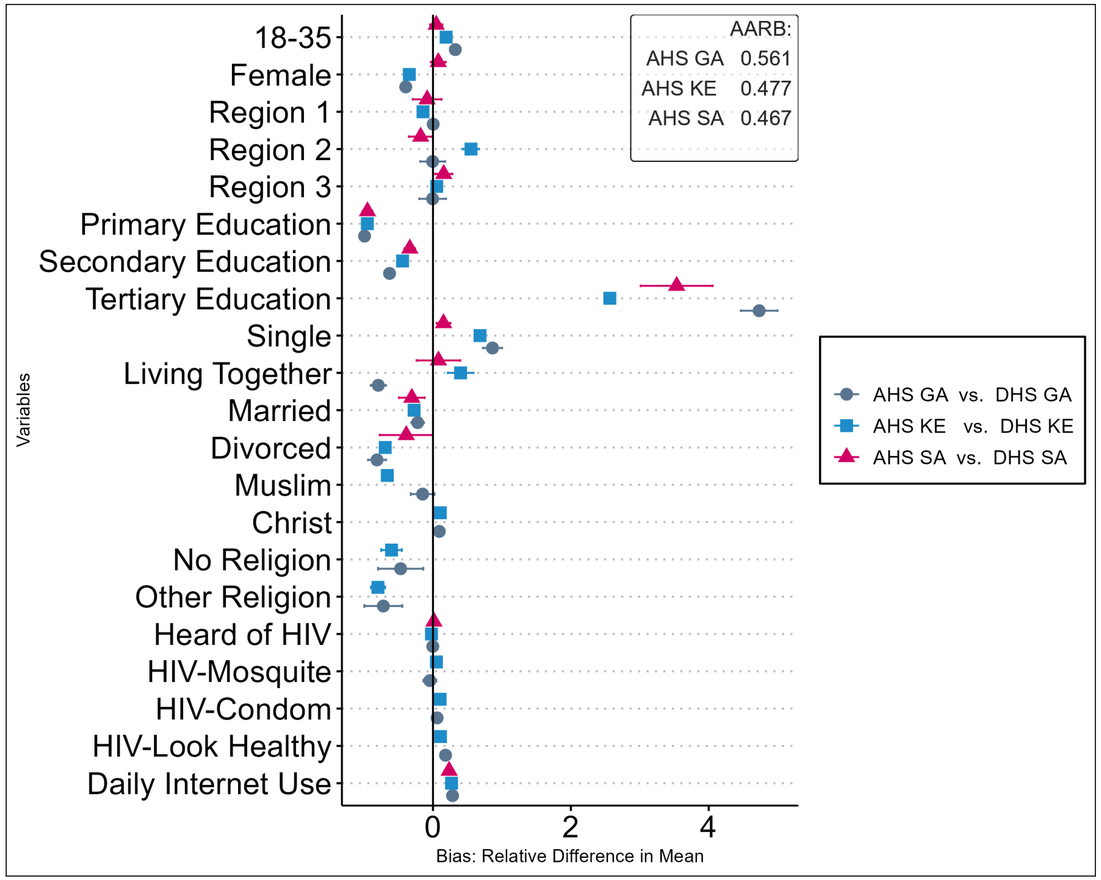
Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```

Analysis accounts for sampling design (e.g., weights and strata)

Alternatives: No weights, different weights in every benchmark.

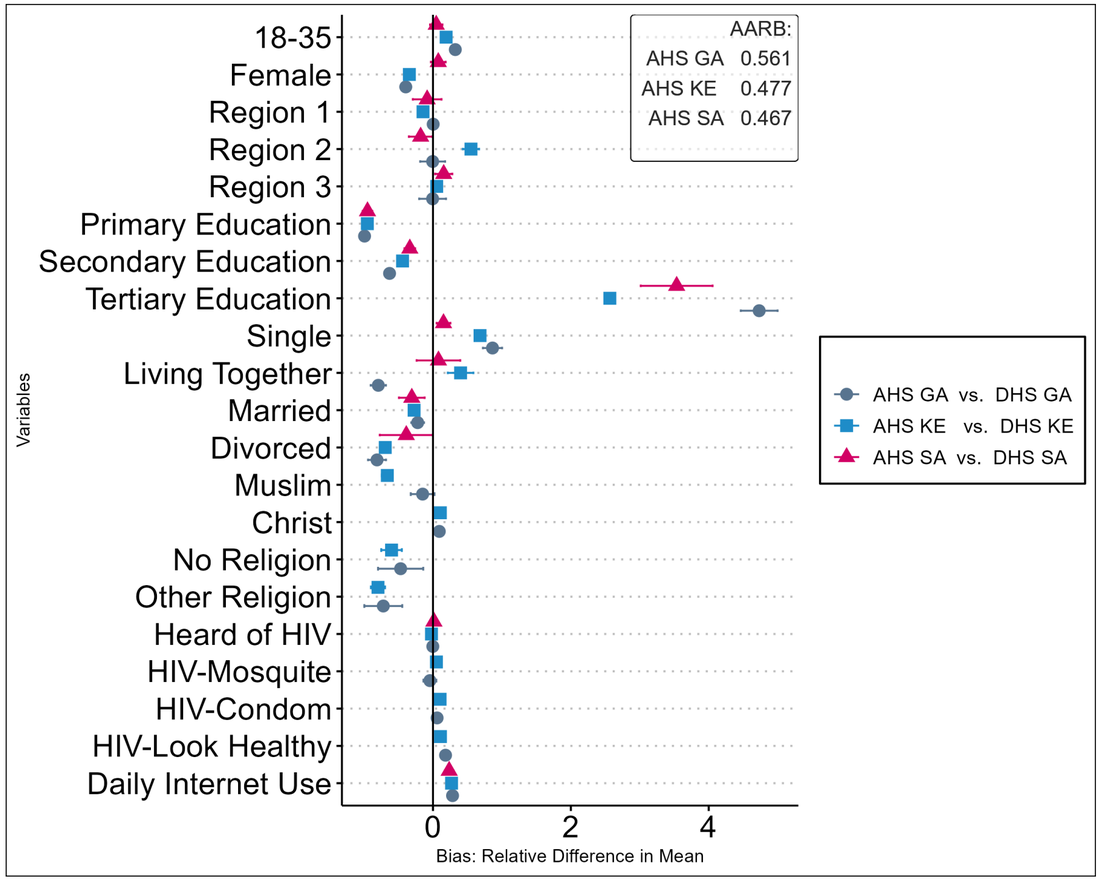


Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```

Variable labels and colors can be customized, then the results are plotted

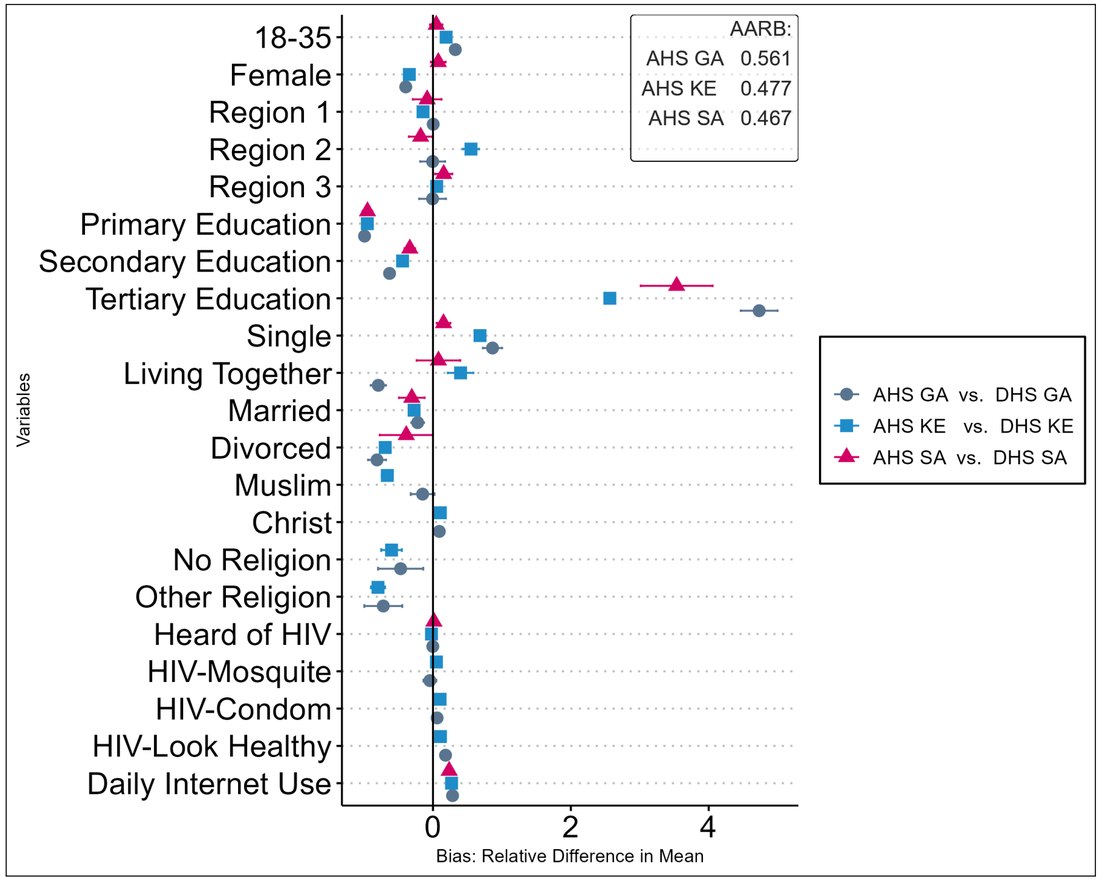


Example Univariate Comparison

Univariate Plot

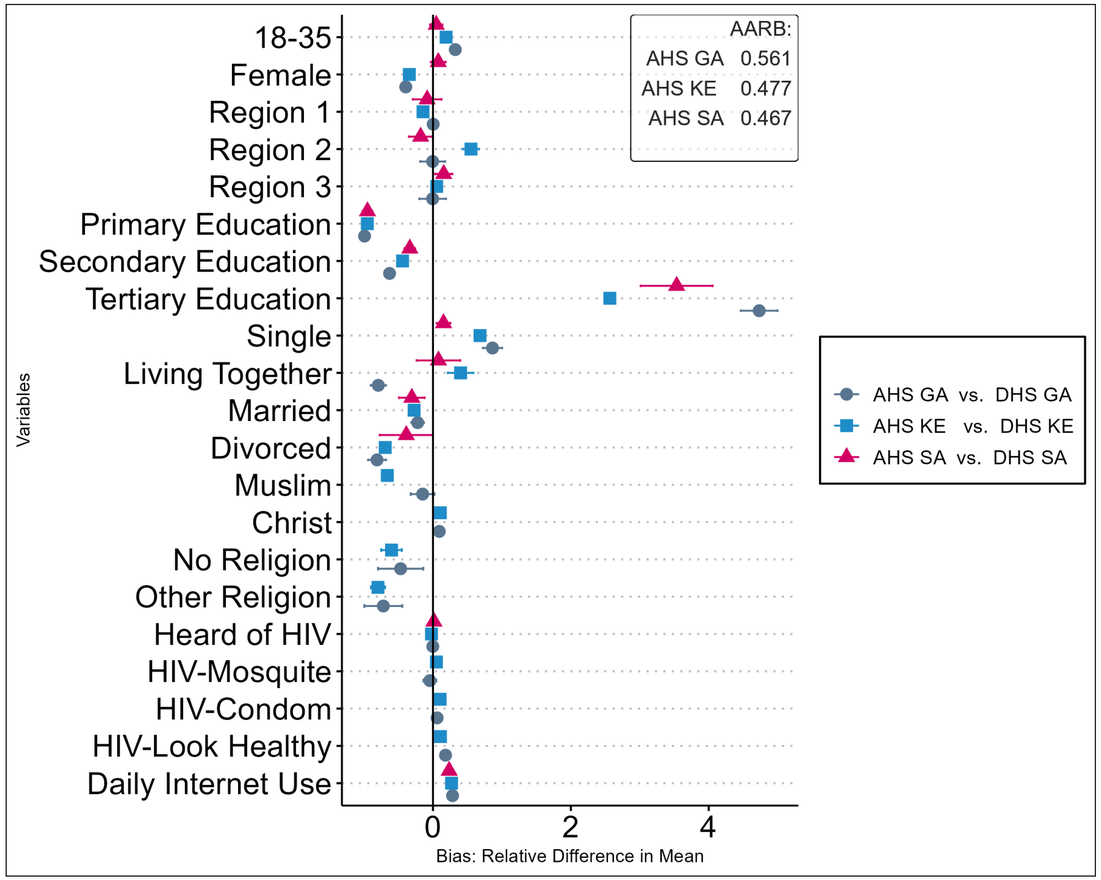
```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```

We put the results of uni_compare in the plot function to get a plot.



Example Univariate Comparison

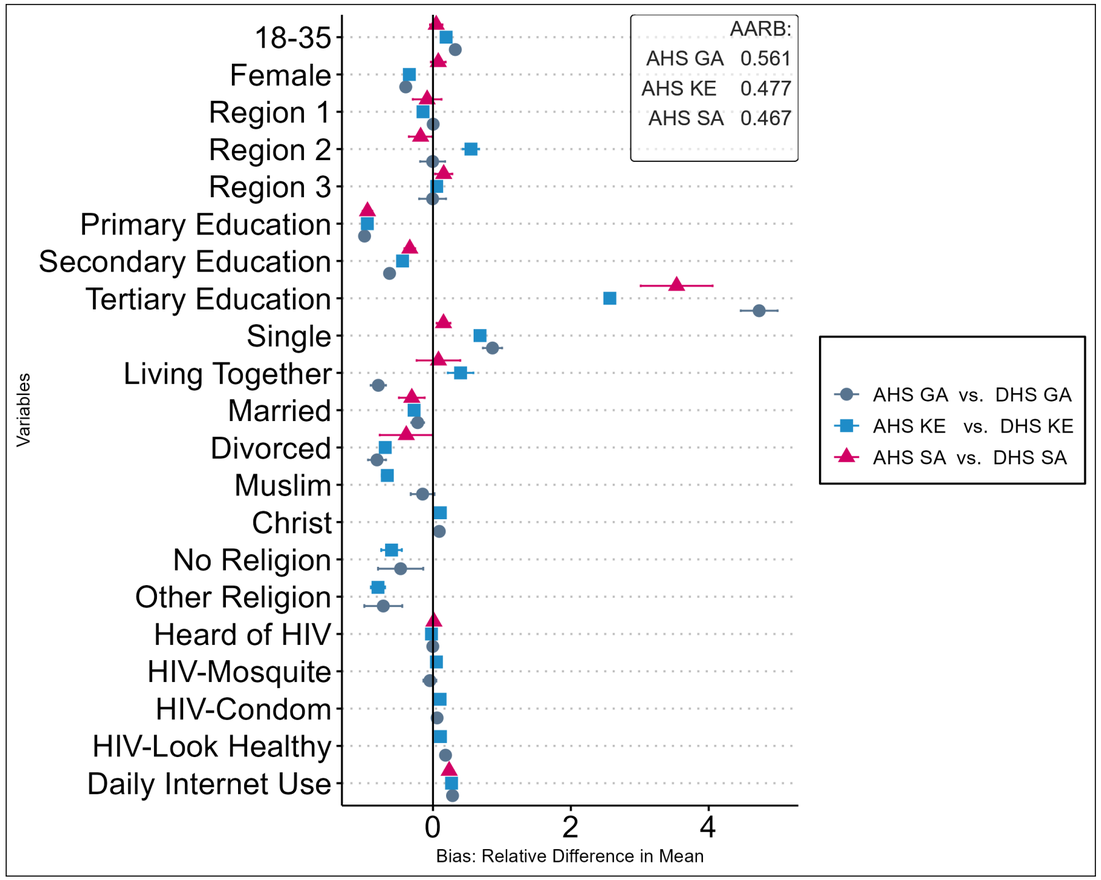
Univariate Plot



Univariate Plot

Results

- Most estimates are rather unbiased
- Tertiary-educated participants are strongly overrepresented
- The surveys are similarly biased on average.



Example Univariate Comparison

Univariate Table

```
univar_table <- uni_compare_table(bias_AHS, ci_line=F)


univar_kable_ahs <- univar_table |>
  kable(caption = "Bias: Difference in Proportions Between AHS and DHS",
        align = c("l","c","c"),
        booktabs = TRUE) |>
  row_spec(0, bold=TRUE) |>
  footnote(
    general_title = "Note.",
    general = "Bias; difference in proportions for dichotome
    variables; confidence intervals (in\n paranthesis) where
    measured by bootstrapping with 2000 iterations; N varries
    between\n variables and surveys due to missing information.",
    footnote_as_chunk = T
  ) |>
  kable_classic(full_width =F, position="center", font_size=14) |>
  row_spec(21, extra_css = "border-bottom: 1.5px solid;")

univar_kable_ahs
```

Bias: Difference in Proportions Between AHS and DHS

Variables	AHS GA	AHS KE	AHS SA
18-35	0.324 (0.267, 0.381)	0.191 (0.168, 0.214)	0.047 (-0.042, 0.137)
Female	-0.395 (-0.474, -0.316)	-0.343 (-0.378, -0.308)	0.077 (-0.030, 0.184)
Region 1	0.003 (-0.062, 0.068)	-0.145 (-0.177, -0.113)	-0.084 (-0.292, 0.123)
Region 2	-0.005 (-0.186, 0.176)	0.551 (0.434, 0.669)	-0.179 (-0.349, -0.009)
Region 3	-0.005 (-0.200, 0.189)	0.054 (-0.011, 0.119)	0.156 (0.031, 0.280)
Primary Education	-0.993 (-1.007, -0.978)	-0.952 (-0.965, -0.939)	-0.950 (-1.020, -0.880)
Secondary Education	-0.630 (-0.691, -0.568)	-0.441 (-0.483, -0.400)	-0.336 (-0.413, -0.259)
Tertiary Education	4.732 (4.465, 4.999)	2.566 (2.489, 2.642)	3.535 (3.014, 4.056)
Single	0.864 (0.724, 1.003)	0.682 (0.618, 0.747)	0.153 (0.053, 0.253)
Living Together	-0.793 (-0.904, -0.683)	0.401 (0.215, 0.586)	0.079 (-0.237, 0.395)
Married	-0.222 (-0.314, -0.130)	-0.274 (-0.308, -0.239)	-0.307 (-0.493, -0.121)
Divorced	-0.812 (-0.942, -0.681)	-0.692 (-0.754, -0.631)	-0.387 (-0.772, -0.002)
Muslim	-0.150 (-0.322, 0.022)	-0.662 (-0.742, -0.583)	NA
Christ	0.092 (0.043, 0.141)	0.107 (0.098, 0.116)	NA
No Religion	-0.469 (-0.797, -0.141)	-0.601 (-0.750, -0.451)	NA
Other Religion	-0.721 (-0.996, -0.446)	-0.798 (-0.900, -0.695)	NA
Heard of HIV	-0.003 (-0.020, 0.015)	-0.022 (-0.028, -0.016)	0.014 (-0.004, 0.031)
HIV-Mosquite	-0.045 (-0.138, 0.048)	0.050 (0.040, 0.060)	NA
HIV-Condom	0.060 (>0.000, 0.119)	0.103 (0.093, 0.113)	NA
HIV-Look Healthy	0.182 (0.118, 0.247)	0.108 (0.100, 0.117)	NA
Daily Internet Use	0.284 (0.253, 0.314)	0.270 (0.254, 0.285)	0.235 (0.172, 0.298)
Average Error	0.561	0.477	0.467
RANK	3	2	1
N	111 - 494	1890 - 2552	275

Note. Bias; difference in proportions for dichotome variables; confidence intervals (in paranthesis) where measured by bootstrapping with 2000 iterations; N varries between variables and surveys due to missing information.

Example Univariate Comparison

Univariate Table

```
univar_table <- uni_compare_table(bias_AHS, ci_line=F)


univar_kable_ahs <- univar_table |>
  kable(caption = "Bias: Difference in Proportions Between AHS and DHS",
        align = c("l","c","c"),
        booktabs = TRUE) |>
  row_spec(0, bold=TRUE) |>
  footnote(
    general_title = "Note.",
    general = "Bias; difference in proportions for dichotome
    variables; confidence intervals (in\n paranthesis) where
    measured by bootstrapping with 2000 iterations; N varries
    between\n variables and surveys due to missing information.",
    footnote_as_chunk = T
  ) |>
  kable_classic(full_width =F, position="center", font_size=14) |>
  row_spec(21, extra_css = "border-bottom: 1.5px solid;")

univar_kable_ahs
```

ci_line = F displays confidence intervals behind the bias
ci_line = T would display them in an additional line beneath

Bias: Difference in Proportions Between AHS and DHS

Variables	AHS GA	AHS KE	AHS SA
18-35	0.324 (0.267, 0.381)	0.191 (0.168, 0.214)	0.047 (-0.042, 0.137)
Female	-0.395 (-0.474, -0.316)	-0.343 (-0.378, -0.308)	0.077 (-0.030, 0.184)
Region 1	0.003 (-0.062, 0.068)	-0.145 (-0.177, -0.113)	-0.084 (-0.292, 0.123)
Region 2	-0.005 (-0.186, 0.176)	0.551 (0.434, 0.669)	-0.179 (-0.349, -0.009)
Region 3	-0.005 (-0.200, 0.189)	0.054 (-0.011, 0.119)	0.156 (0.031, 0.280)
Primary Education	-0.993 (-1.007, -0.978)	-0.952 (-0.965, -0.939)	-0.950 (-1.020, -0.880)
Secondary Education	-0.630 (-0.691, -0.568)	-0.441 (-0.483, -0.400)	-0.336 (-0.413, -0.259)
Tertiary Education	4.732 (4.465, 4.999)	2.566 (2.489, 2.642)	3.535 (3.014, 4.056)
Single	0.864 (0.724, 1.003)	0.682 (0.618, 0.747)	0.153 (0.053, 0.253)
Living Together	-0.793 (-0.904, -0.683)	0.401 (0.215, 0.586)	0.079 (-0.237, 0.395)
Married	-0.222 (-0.314, -0.130)	-0.274 (-0.308, -0.239)	-0.307 (-0.493, -0.121)
Divorced	-0.812 (-0.942, -0.681)	-0.692 (-0.754, -0.631)	-0.387 (-0.772, -0.002)
Muslim	-0.150 (-0.322, 0.022)	-0.662 (-0.742, -0.583)	NA
Christ	0.092 (0.043, 0.141)	0.107 (0.098, 0.116)	NA
No Religion	-0.469 (-0.797, -0.141)	-0.601 (-0.750, -0.451)	NA
Other Religion	-0.721 (-0.996, -0.446)	-0.798 (-0.900, -0.695)	NA
Heard of HIV	-0.003 (-0.020, 0.015)	-0.022 (-0.028, -0.016)	0.014 (-0.004, 0.031)
HIV-Mosquite	-0.045 (-0.138, 0.048)	0.050 (0.040, 0.060)	NA
HIV-Condom	0.060 (>0.000, 0.119)	0.103 (0.093, 0.113)	NA
HIV-Look Healthy	0.182 (0.118, 0.247)	0.108 (0.100, 0.117)	NA
Daily Internet Use	0.284 (0.253, 0.314)	0.270 (0.254, 0.285)	0.235 (0.172, 0.298)
Average Error	0.561	0.477	0.467
RANK	3	2	1
N	111 - 494	1890 - 2552	275

Note. Bias; difference in proportions for dichotome variables; confidence intervals (in paranthesis) where measured by bootstrapping with 2000 iterations; N varries between variables and surveys due to missing information.

GESIS Leibniz Institute for the Social Sciences

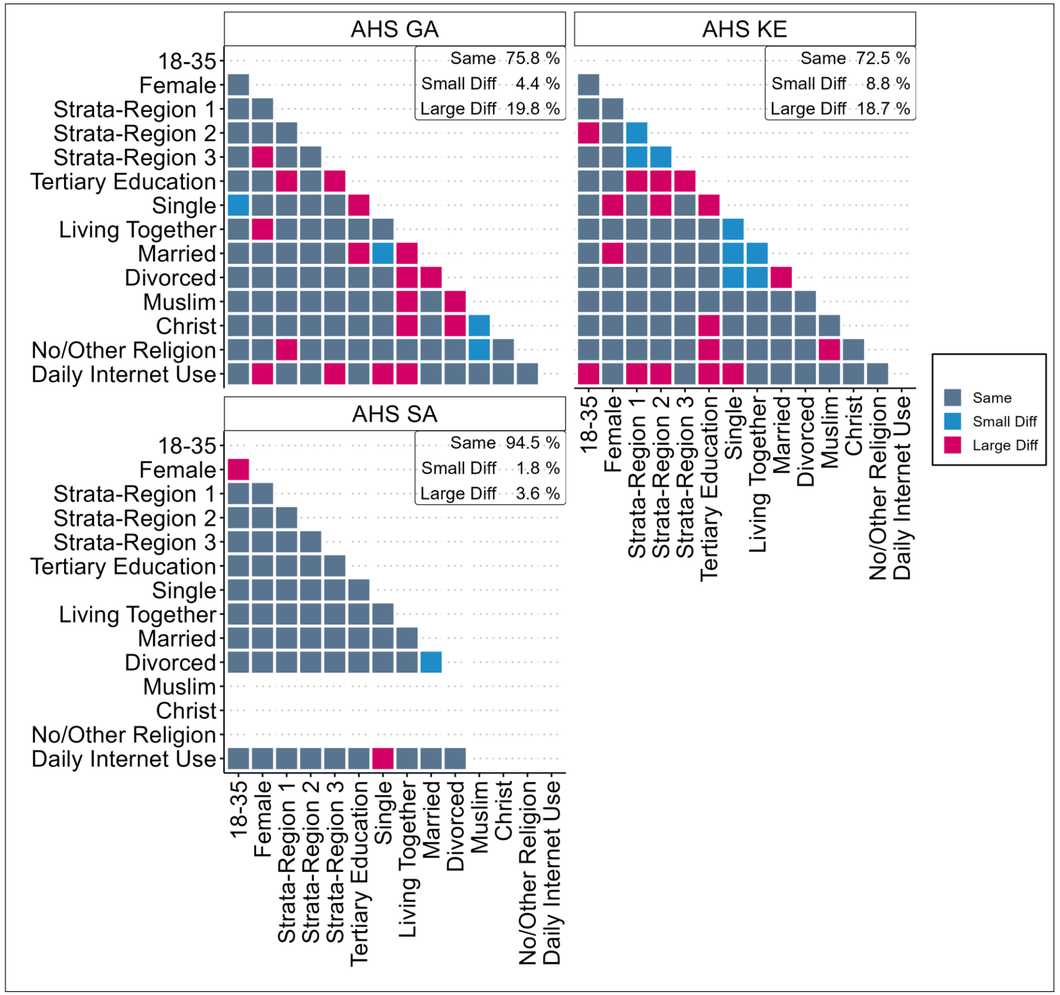
Example Bivariate Comparison

Leibniz Association

Example Bivariate Comparison

Bivariate Plot

```
biv_AHS_weighted_DHS <- biv_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                                    benchmarks = c("DHS_Ga",
                                                   "DHS_Ke",
                                                   "DHS_Sa"),
                                    variables = variables2,
                                    plots_label = c("AHS GA",
                                                    "AHS KE",
                                                    "AHS SA"),
                                    p_adjust="bonferroni",
                                    weight_bench = c("weight"),
                                    id_bench = c("cluster"),
                                    strata_bench = c("strata"),
                                    varlabels = varlabels2, nboots =2000,
                                    colors = colors)

biv_plot<-plot_biv_compare(biv_AHS_weighted_DHS,ncol_facet = 2)
```

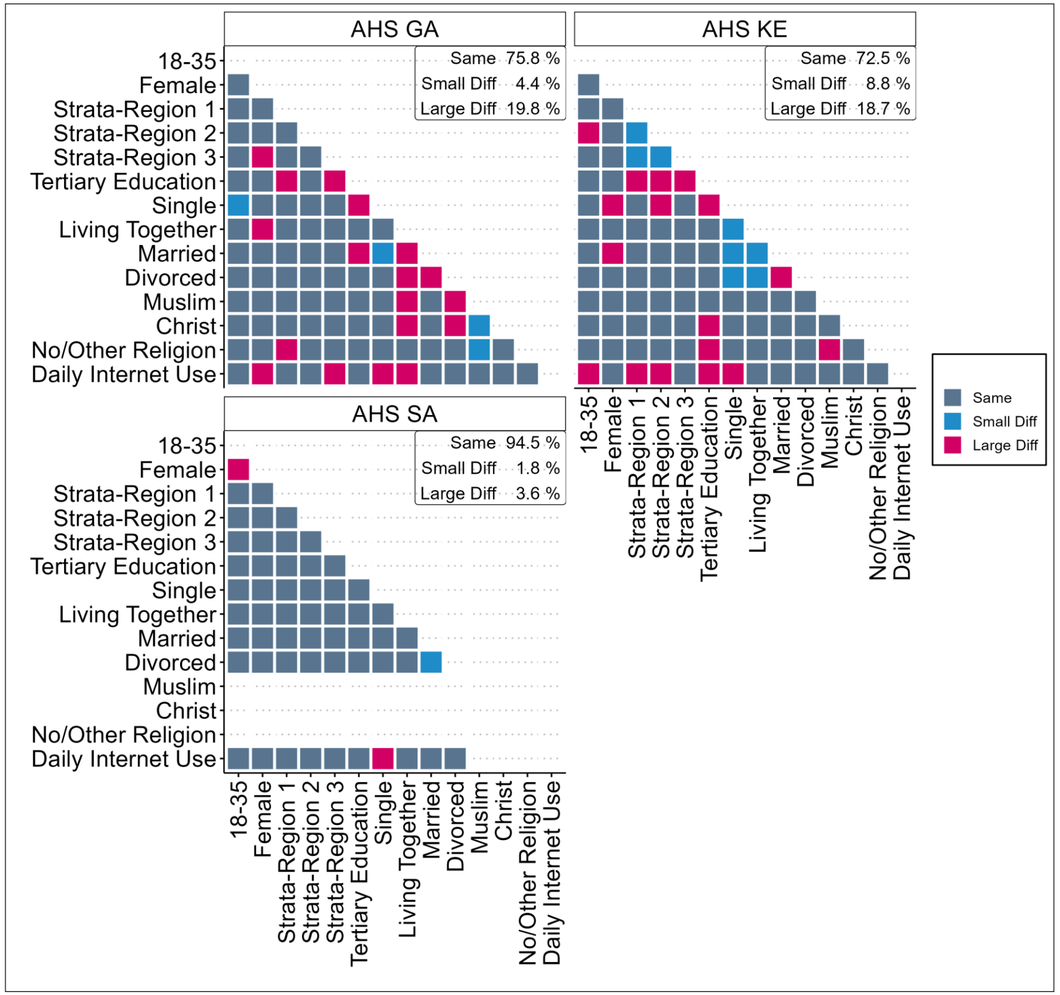


Example Bivariate Comparison

Bivariate Plot

```
biv_AHS_weighted_DHS <- biv_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                                    benchmarks = c("DHS_Ga",
                                                   "DHS_Ke",
                                                   "DHS_Sa"),
                                    variables = variables2,
                                    plots_label = c("AHS GA",
                                                    "AHS KE",
                                                    "AHS SA"),
                                    p_adjust="bonferroni",
                                    weight_bench = c("weight"),
                                    id_bench = c("cluster"),
                                    strata_bench = c("strata"),
                                    varlabels = varlabels2, nboots =2000,
                                    colors = colors)

biv_plot<-plot_biv_compare(biv_AHS_weighted_DHS,ncol_facet = 2)
```

Similar inputs to the function as before.

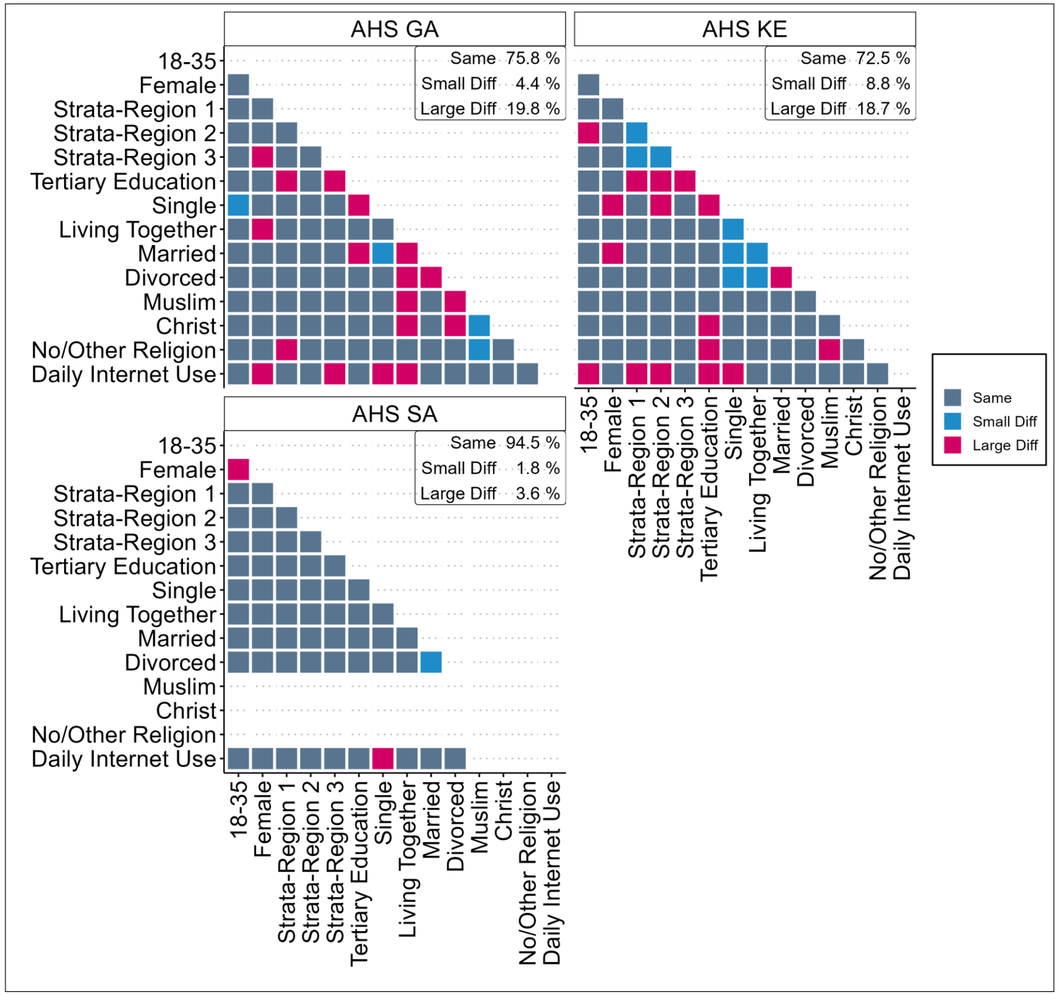


Example Bivariate Comparison

Bivariate Plot

```
biv_AHS_weighted_DHS <- biv_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                                    benchmarks = c("DHS_Ga",
                                                   "DHS_Ke",
                                                   "DHS_Sa"),
                                    variables = variables2,
                                    plots_label = c("AHS GA",
                                                    "AHS KE",
                                                    "AHS SA"),
                                    p_adjust="bonferroni",
                                    weight_bench = c("weight"),
                                    id_bench = c("cluster"),
                                    strata_bench = c("strata"),
                                    varlabels = varlabels2, nboots =2000,
                                    colors = colors)

biv_plot<-plot_biv_compare(biv_AHS_weighted_DHS,ncol_facet = 2)
```

1. Pearson's r is calculated for every pair of variables in surveys and benchmarks.
2. They are compared against each other to evaluate the proportion of significantly different correlations.



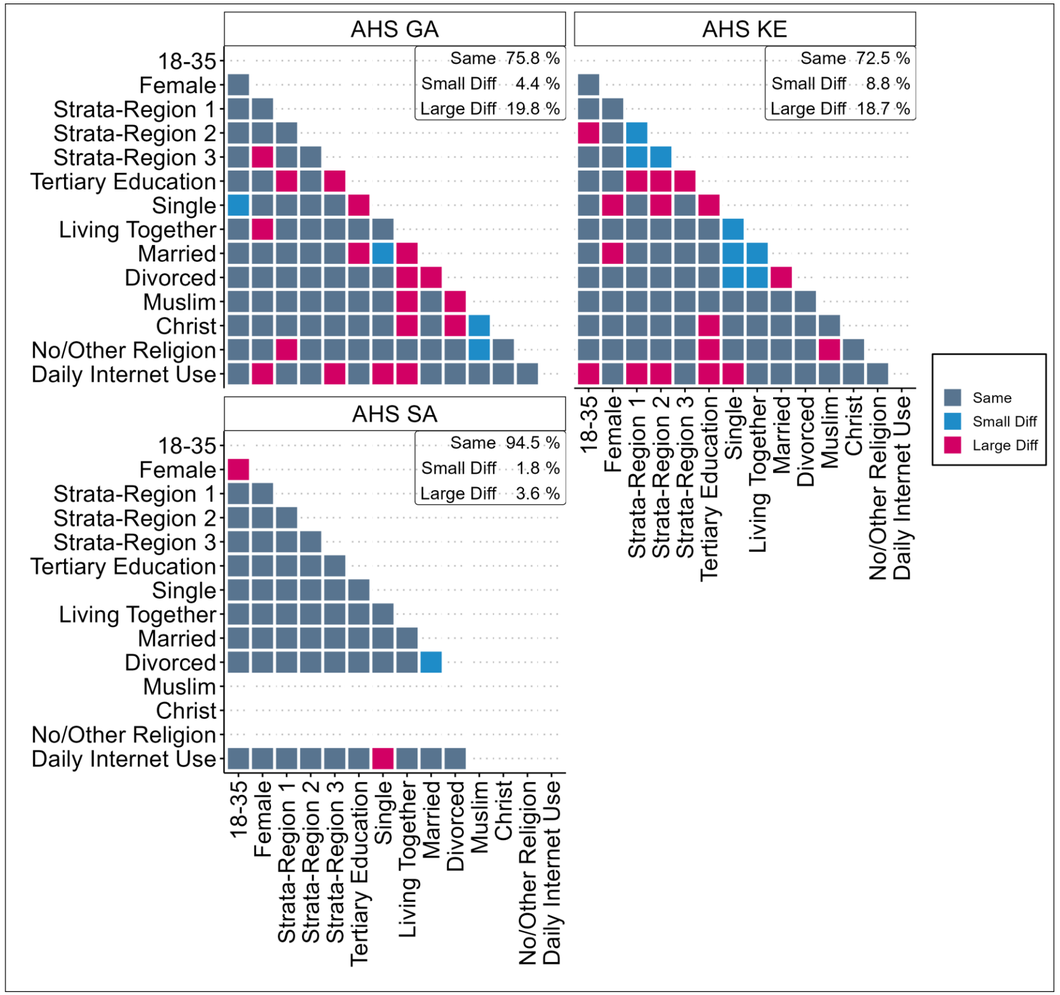
Example Bivariate Comparison

Bivariate Plot

```
biv_AHS_weighted_DHS <- biv_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                                    benchmarks = c("DHS_Ga",
                                                   "DHS_Ke",
                                                   "DHS_Sa"),
                                    variables = variables2,
                                    plots_label = c("AHS GA",
                                                    "AHS KE",
                                                    "AHS SA"),
                                    p_adjust="bonferroni",
                                    weight_bench = c("weight"),
                                    id_bench = c("cluster"),
                                    strata_bench = c("strata"),
                                    varlabels = varlabels2, nboots =2000,
                                    colors = colors)

biv_plot<-plot_biv_compare(biv_AHS_weighted_DHS,ncol_facet = 2)
```

Same = Pearson's r is not significantly* different between surveys or no significant correlation in survey or benchmark.
Small Diff = Pearson's r is significantly* different between survey and benchmark & one of them is significantly* different from 0.
Large Diff = "Small Diff" + one Pearson's r is double the size of the other, or they differ in direction.

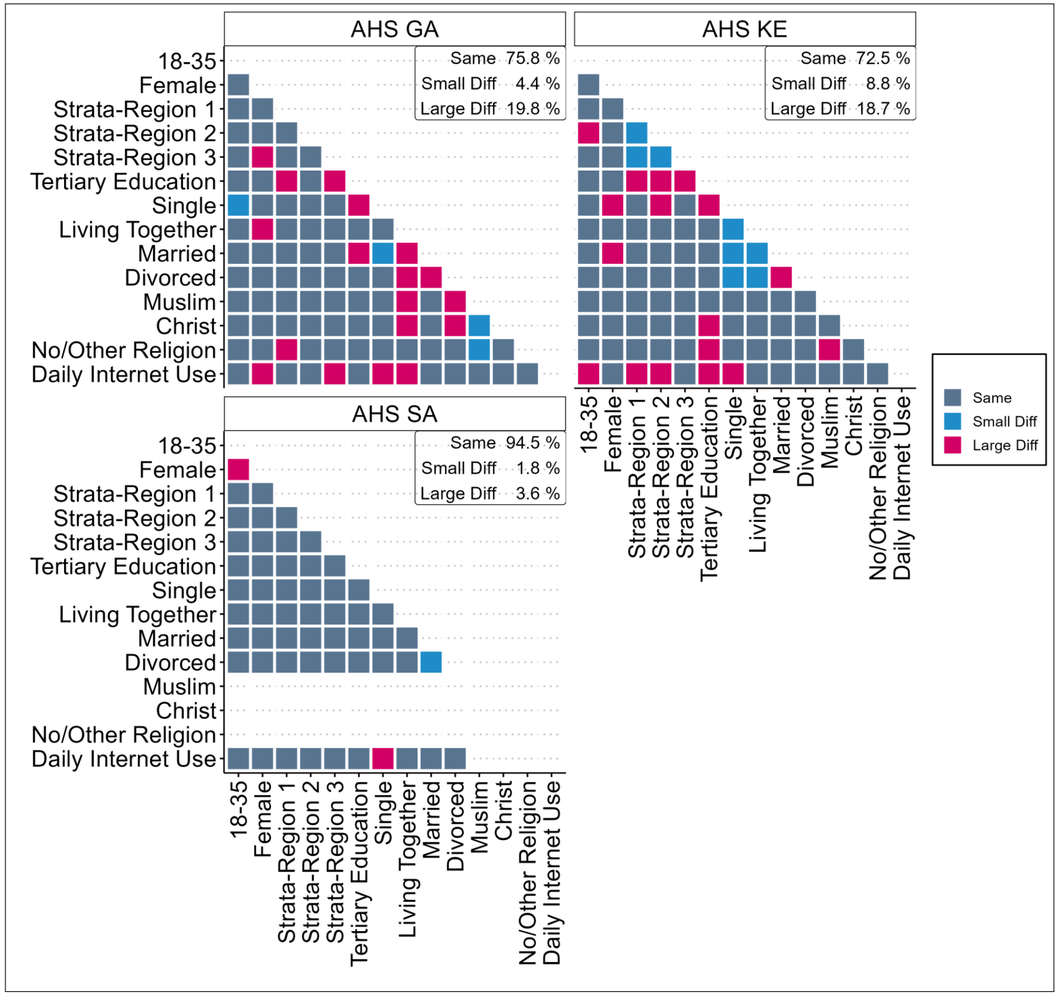
*p< 0.05



Bivariate Plot

Results

- In Ghana and Kenya, some bivariate estimates were biased, while in South Africa, only a few were biased.
- Most of the time when the estimates were significantly different from the benchmark, they were also strongly different from the DHS



GESIS Leibniz Institute for the Social Sciences

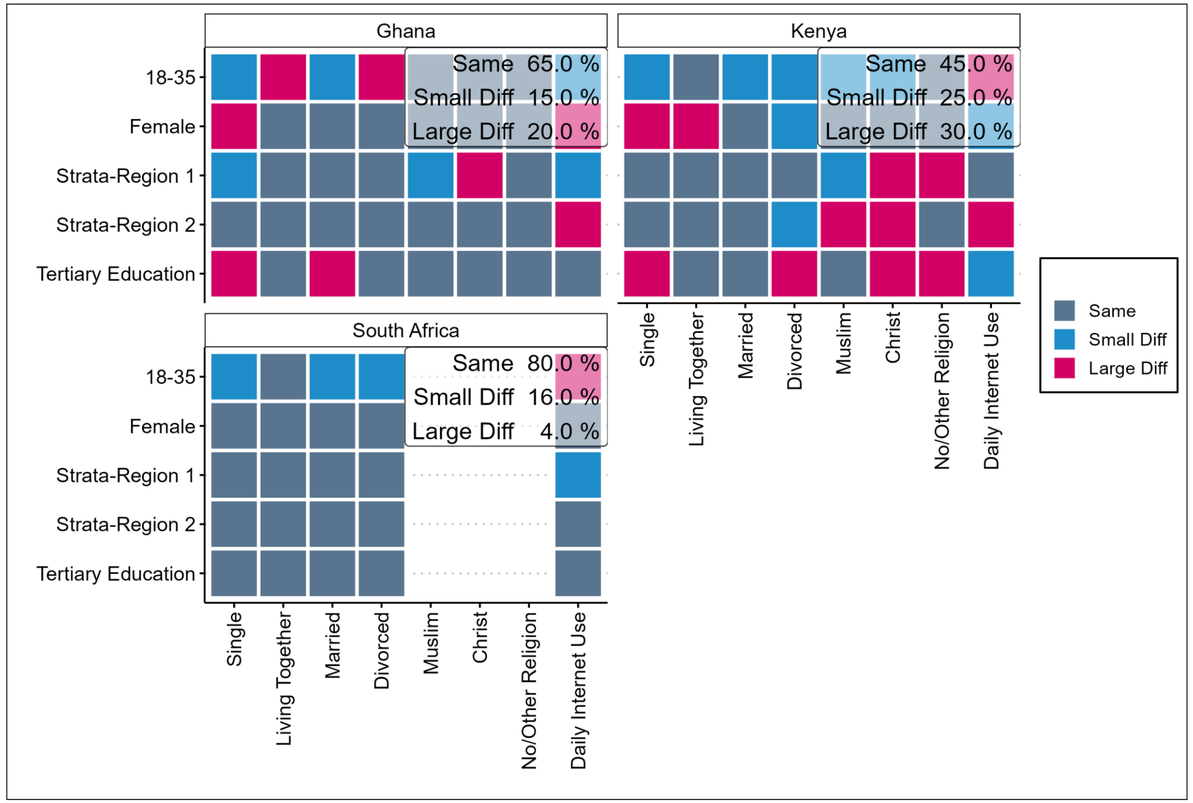
Example Multivariate Comparison

Leibniz Association

Example Multivariate Comparison

Example Multivariate Plot

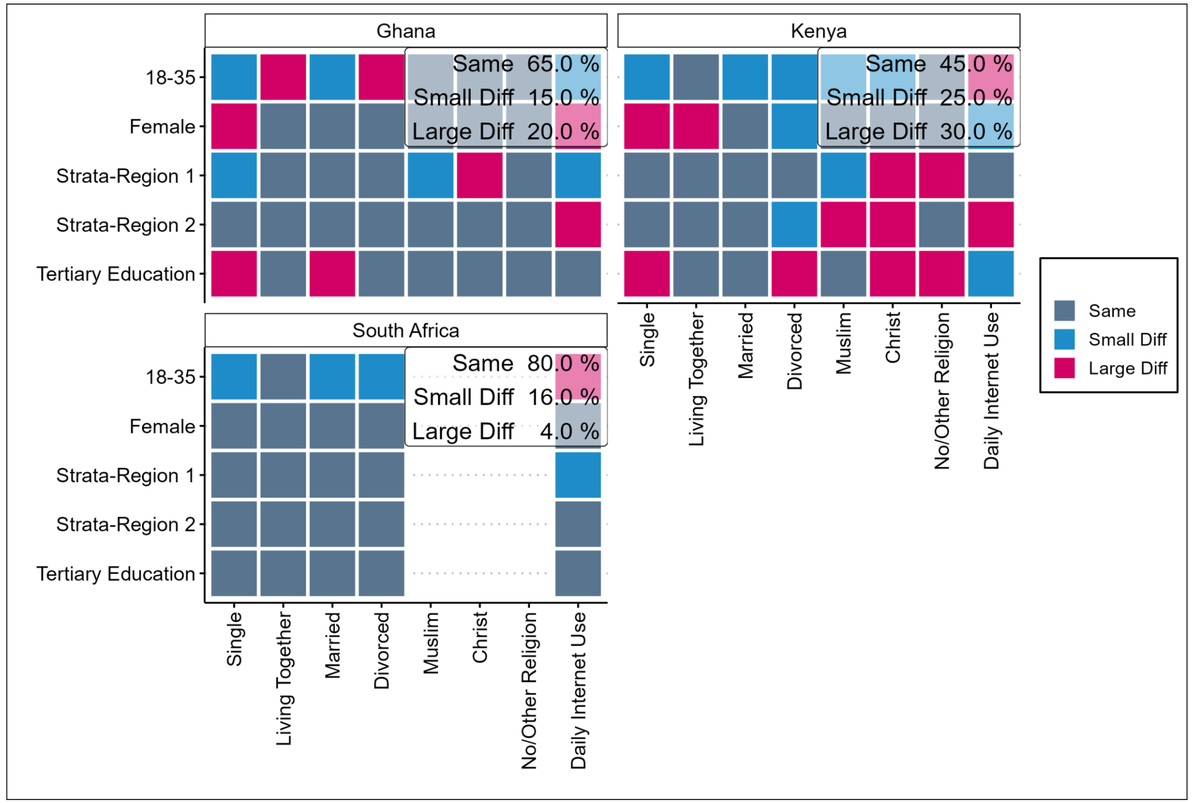
```
multi_data_GA<-multi_compare(df = AHS_GA, benchmark = DHS_Ga,
                             dependent = dependent ,independent = independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_KE<-multi_compare(df = AHS_KE, benchmark = DHS_Ke,
                             dependent =dependent ,independent =independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_SA<-multi_compare(df = AHS_SA, benchmark = DHS_Sa,
                             dependent =dependent ,independent =independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")


multi_plots_Africa2<-
  plot_multi_compare(multi_compare_objects = c("multi_data_GA",
                                               "multi_data_KE",
                                               "multi_data_SA"),
                                    perc_diff_transparance = 0.5,
                                    label_x = xlab,
                                    label_y = ylab,
                                    missings_x = F,
                                    colors = colors)
```



Example Multivariate Plot

```
multi_data_GA<-multi_compare(df = AHS_GA, benchmark = DHS_Ga,
                             dependent = dependent ,independent = independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_KE<-multi_compare(df = AHS_KE, benchmark = DHS_Ke,
                             dependent =dependent ,independent =independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_SA<-multi_compare(df = AHS_SA, benchmark = DHS_Sa,
                             dependent =dependent ,independent =independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")


multi_plots_Africa2<-
  plot_multi_compare(multi_compare_objects = c("multi_data_GA",
                                               "multi_data_KE",
                                               "multi_data_SA"),
                                    perc_diff_transparance = 0.5,
                                    label_x = xlab,
                                    label_y = ylab,
                                    missings_x = F,
                                    colors = colors)
```

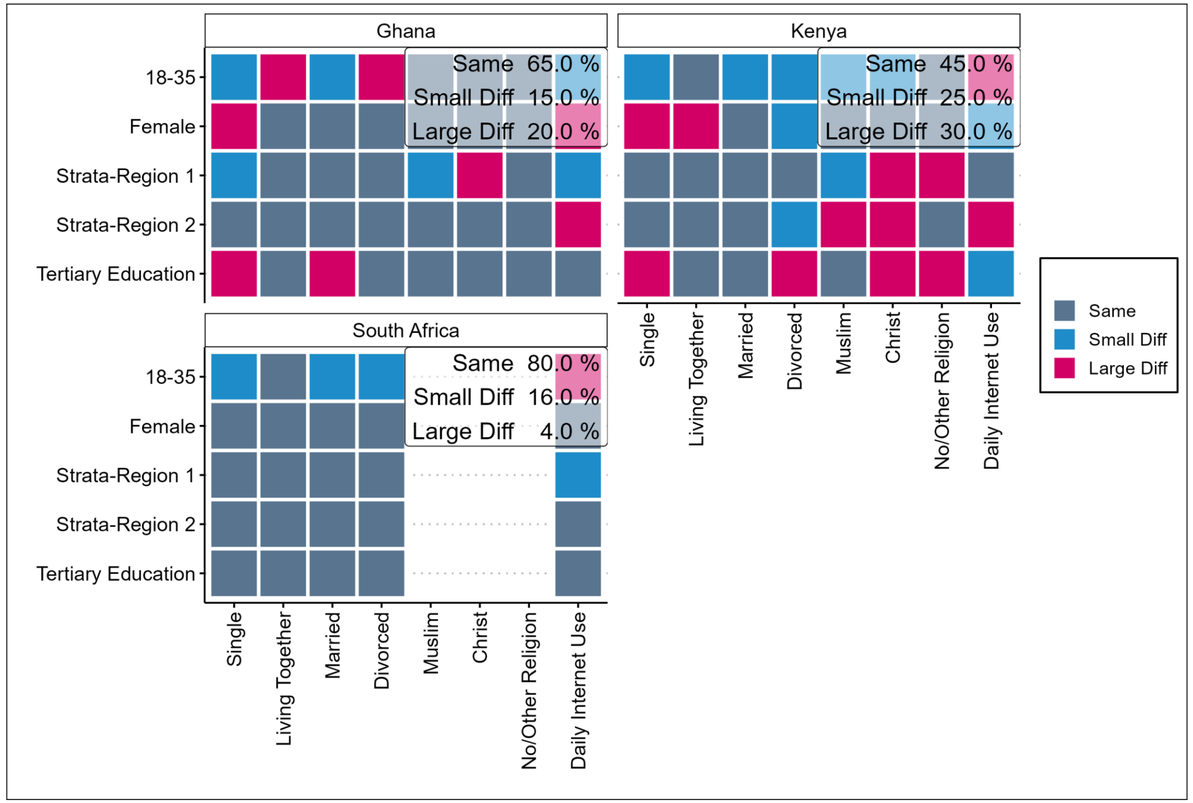
- In the Plot on the right, every model is a column, and every square is a coefficient.
- The color again indicates the difference between benchmark and survey.



Example Multivariate Plot

```
multi_data_GA<-multi_compare(df = AHS_GA, benchmark = DHS_Ga,
                             dependent = dependent ,independent = independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_KE<-multi_compare(df = AHS_KE, benchmark = DHS_Ke,
                             dependent =dependent ,independent =independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_SA<-multi_compare(df = AHS_SA, benchmark = DHS_Sa,
                             dependent =dependent ,independent =independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")


multi_plots_Africa2<-
  plot_multi_compare(multi_compare_objects = c("multi_data_GA",
                                               "multi_data_KE",
                                               "multi_data_SA"),
                                    perc_diff_transparance = 0.5,
                                    label_x = xlab,
                                    label_y = ylab,
                                    missings_x = F,
                                    colors = colors)
```

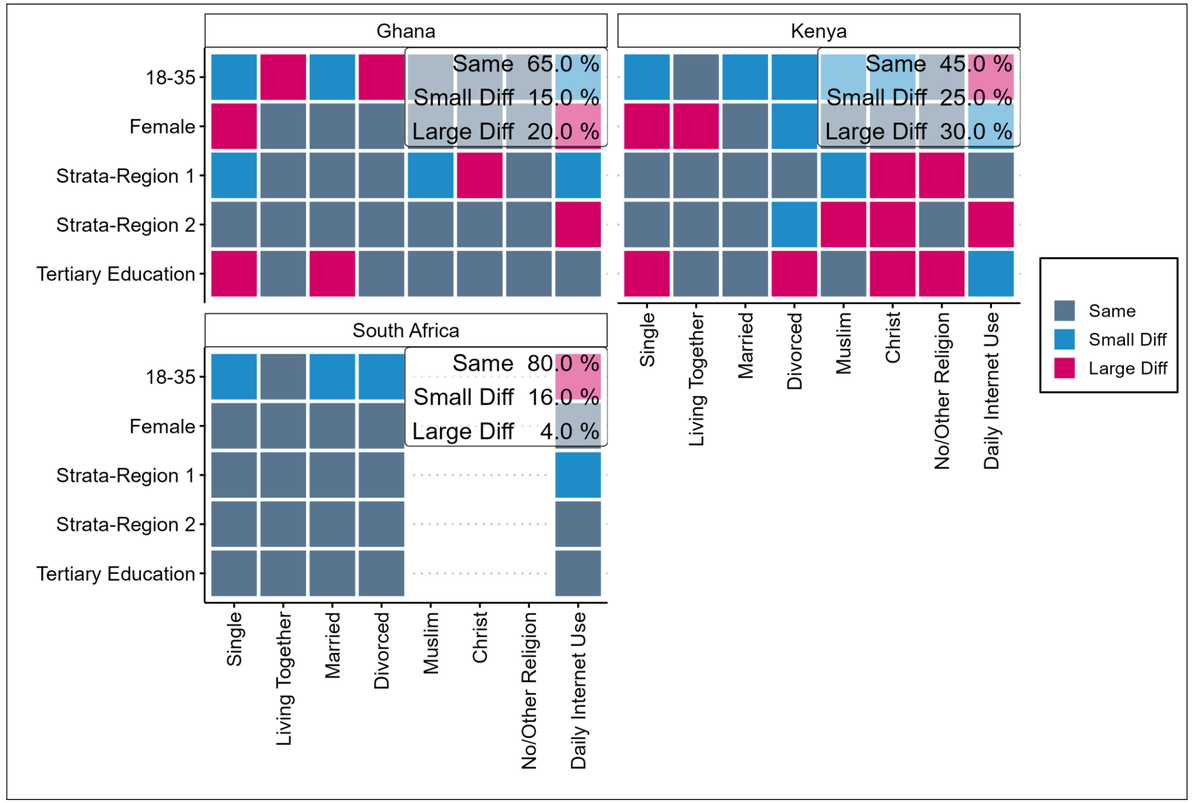
- Models can be specified as a list of dependent and independent variables.
  - A regression will be estimated with every dependent variable using all independent variables.



Example Multivariate Plot

```
independent<- c("young", "female", "strata_reg1", "strata_reg2", "tertiary_educ")
dependent<- c("single", "living_together", "married", "divorced",
              "muslim", "christ","no_rel","internet_daily")
```

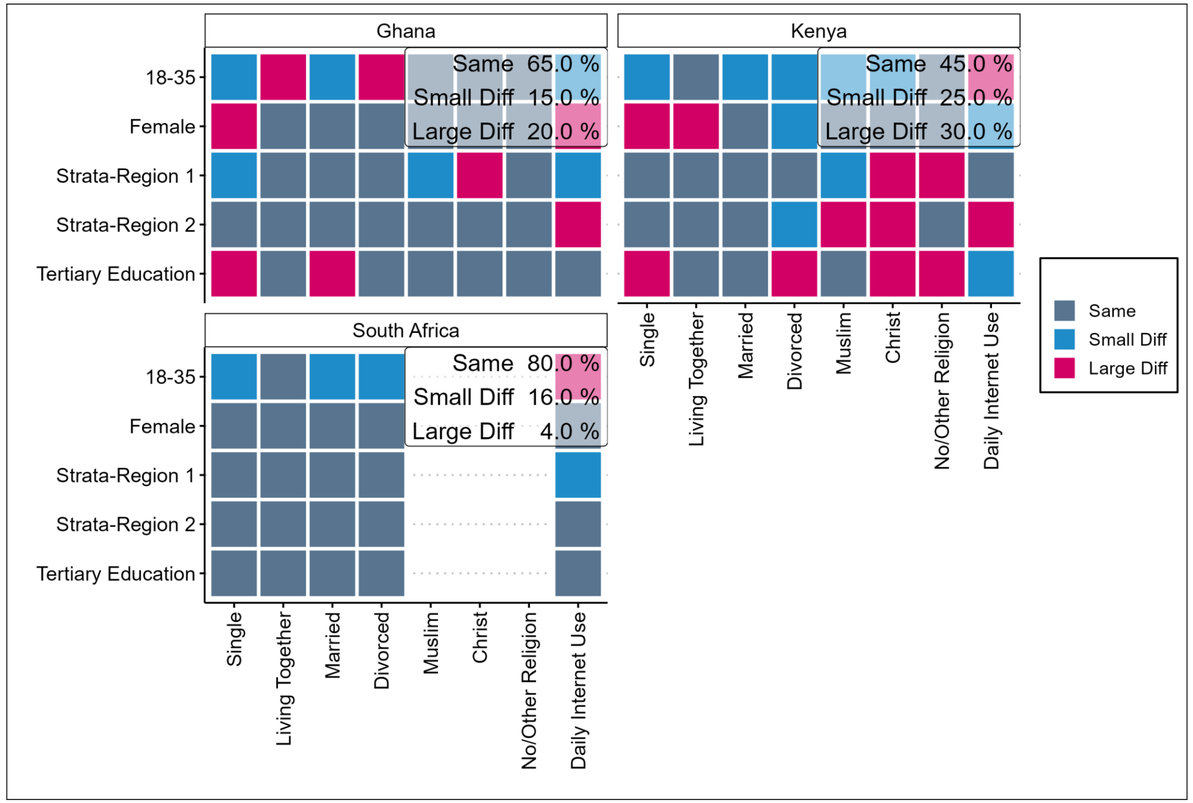
- Models can be specified as a list of dependent and independent variables.
  - A regression will be estimated with every dependent variable using all independent variables.



Example Multivariate Plot

```
multi_data_GA<-multi_compare(df = AHS_GA, benchmark = DHS_Ga,
                             formula_list = form_list,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_KE<-multi_compare(df = AHS_KE, benchmark = DHS_Ke,
                             formula_list = form_list,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_SA<-multi_compare(df = AHS_SA, benchmark = DHS_Sa,
                             formula_list = form_list,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")


multi_plots_Africa2<-
  plot_multi_compare(multi_compare_objects = c("multi_data_GA",
                                               "multi_data_KE",
                                               "multi_data_SA"),
                     perc_diff_transparance = 0.5,
                     label_x = xlab,
                     label_y = ylab,
                     missings_x = F,
                     colors = colors)
```

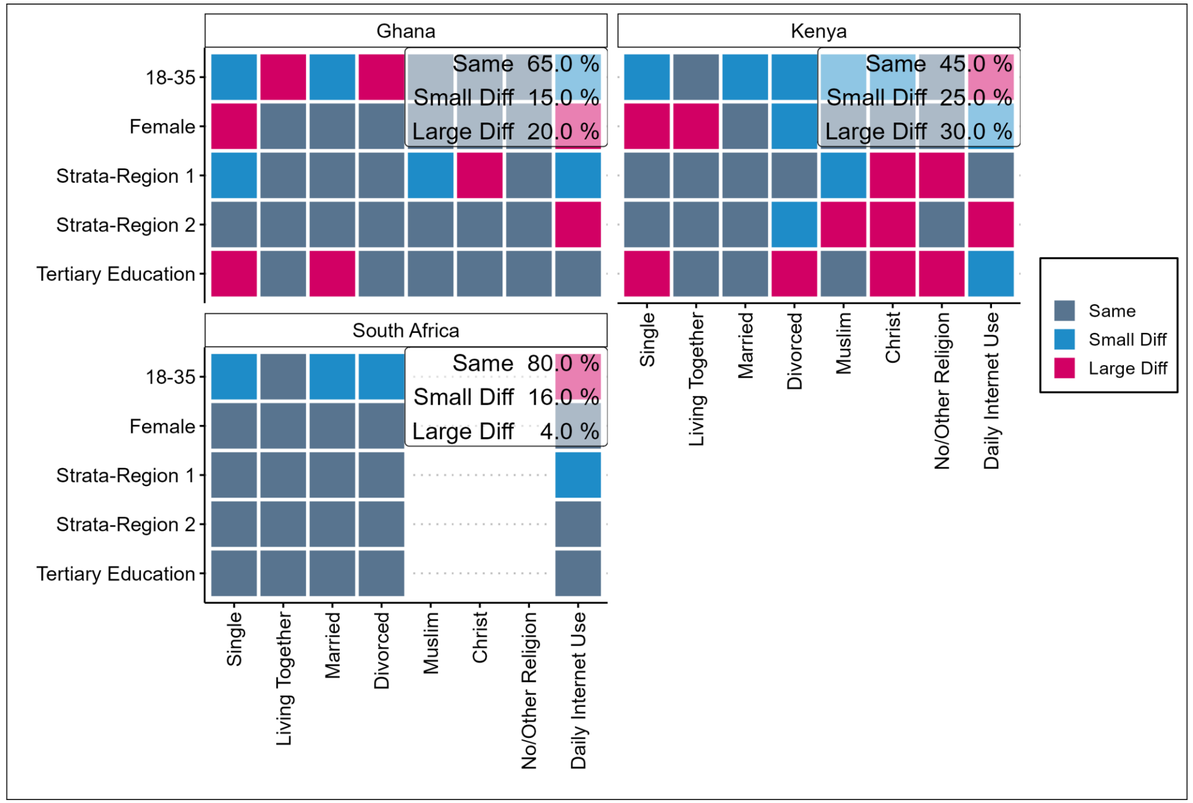
- Models can also be specified as a list of formulas. (i.e. parameter formula_list)
  - A regression will be estimated for every model in the list.



Example Multivariate Plot

```
form_list<-list(formula(single~young+female+strata_reg1+strata_reg2+tertiary_educ),
                formula(living_together~young+female+strata_reg1+strata_reg2+tertiary_educ),
                formula(married~young+female+strata_reg1+strata_reg2+tertiary_educ),
                formula(divorced~young+female+strata_reg1+strata_reg2+tertiary_educ),
                formula(muslim~young+female+strata_reg1+strata_reg2+tertiary_educ),
                formula(christ~young+female+strata_reg1+strata_reg2+tertiary_educ),
                formula(no_rel~young+female+strata_reg1+strata_reg2+tertiary_educ),
                formula(internet_daily~young+female+strata_reg1+strata_reg2+tertiary_educ))
```

- Models can also be specified as a list of formulas. (i.e. parameter formula_list)
  - A regression will be estimated for every model in the list.



GESIS Leibniz Institute for the Social Sciences

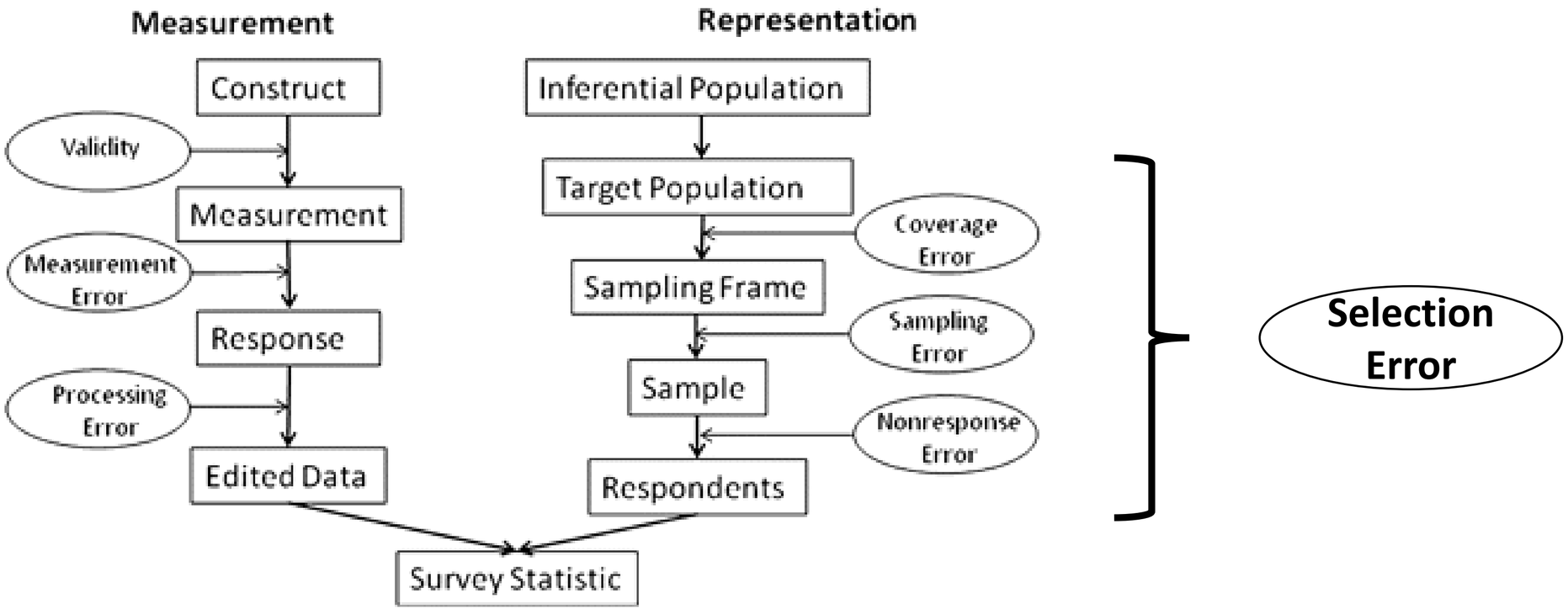
Choosing the right benchmark

Leibniz Association

The right benchmark depends on what you want to analyze

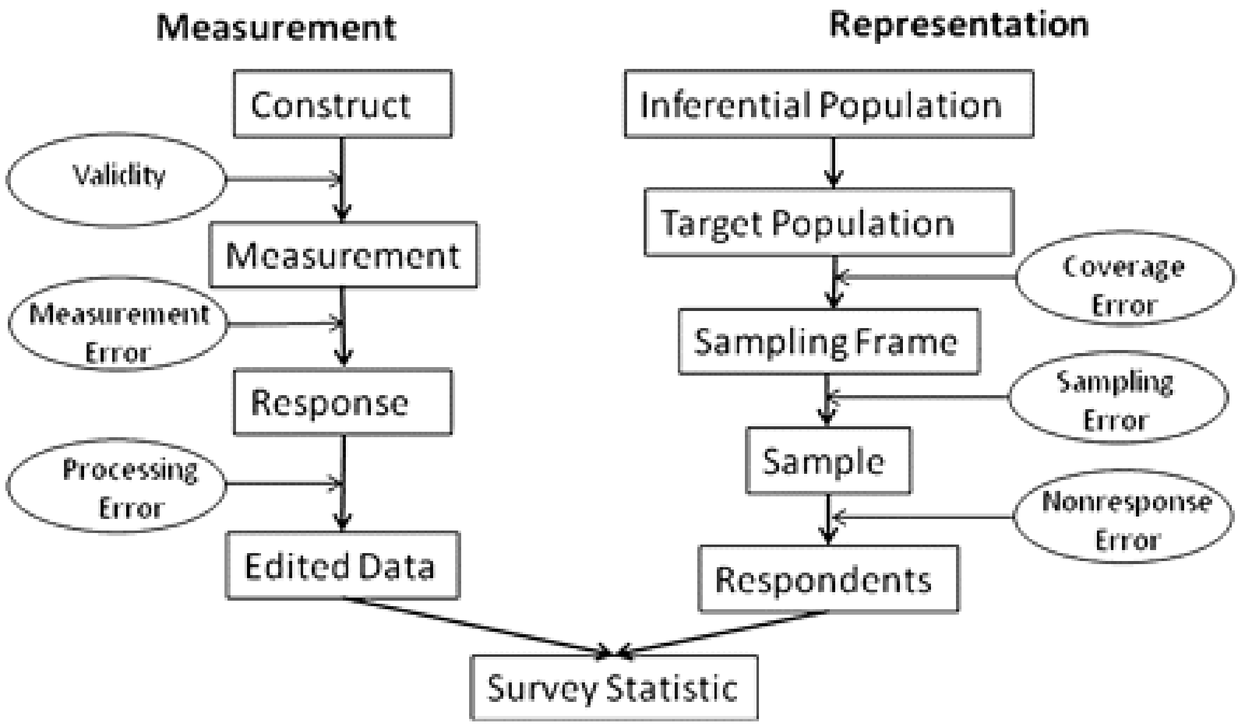
Probability Surveys

Nonprobability Surveys



Total Survey Error Framework (Groves and Lyberg 2010) Reprinted from “Total Survey Error: Past, Present, and Future” by Author R. M. Groves & L. Lyberg, 2010, Public Opinion Quarterly, 74(5), 856. Copyright [2010] by the Oxford University Press. Reprinted with permission.

The right benchmark depends on what you want to analyze



If you want to analyze the overall non-response error, the benchmark one would need should contain information for any sample member.

Total Survey Error Framework (Groves and Lyberg 2010) Reprinted from “Total Survey Error: Past, Present, and Future” by Author R. M. Groves & L. Lyberg, 2010, Public Opinion Quarterly, 74(5), 856. Copyright [2010] by the Oxford University Press. Reprinted with permission.

Analyzing specific types of nonresponse error

- In a mixed-mode survey (e.g., the GESIS Panel), one might want to analyze the bias that would occur when only one mode would be available.

- Who would we have lost when the survey would have only allowed for participating in the online mode? (Lena Rembser @GESIS is currently working on such an analysis)

- In such a situation, one could use the whole survey as a benchmark and compare it against the online respondents.

Quality of the benchmark is very important

- If available and recently conducted:
  - Census and Microcensus
  - Other Official Statistics
  - Sample & Sample Frame Information

- High Quality Probability surveys
  - Low nonresponse,
  - Potentially validated against official statistics benchmarks on other available estimates

GESIS Leibniz Institute for the Social Sciences

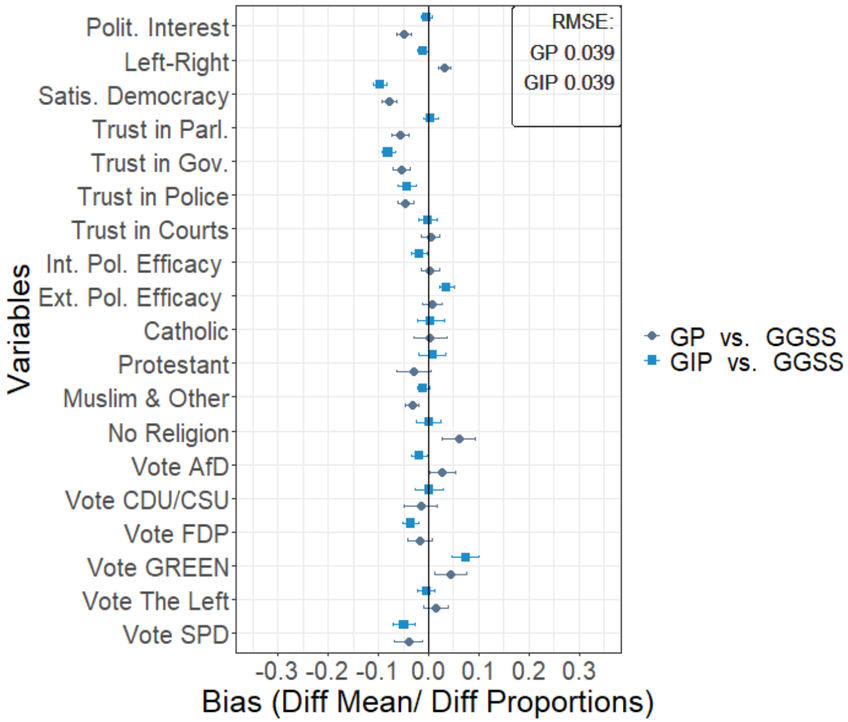
Other Examples

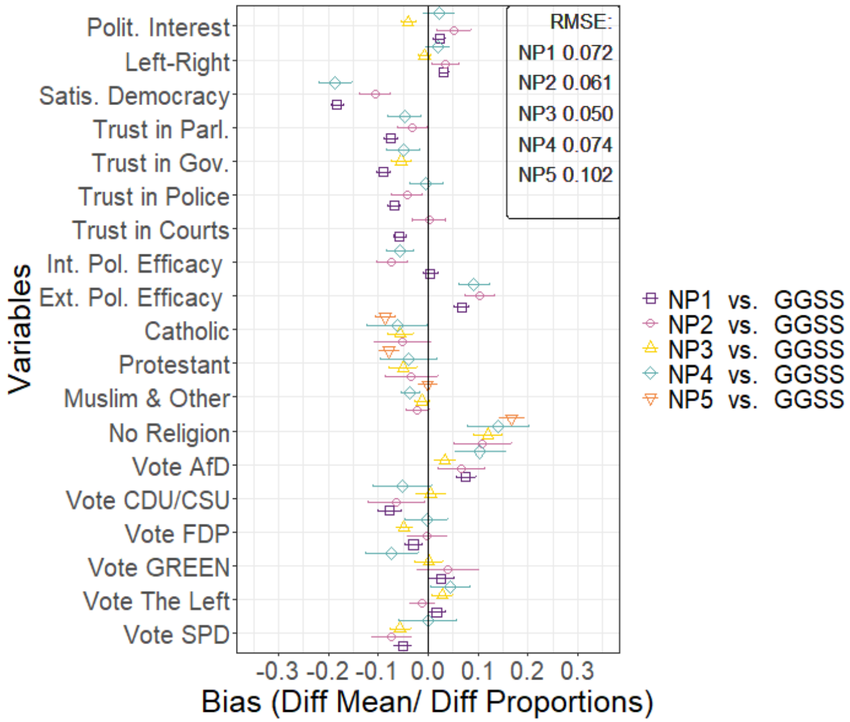
Leibniz Association

Probability and Nonprobability Surveys

- We compared **5 nonprobability surveys** and **3 probability surveys** against population benchmarks.
- 3 nonprobability surveys came from the **same panel**.
- Benchmark was the **German Microcensus** for demographic variables.
- Additionally, we used the best performing probability survey as benchmark for **non-demographics**.

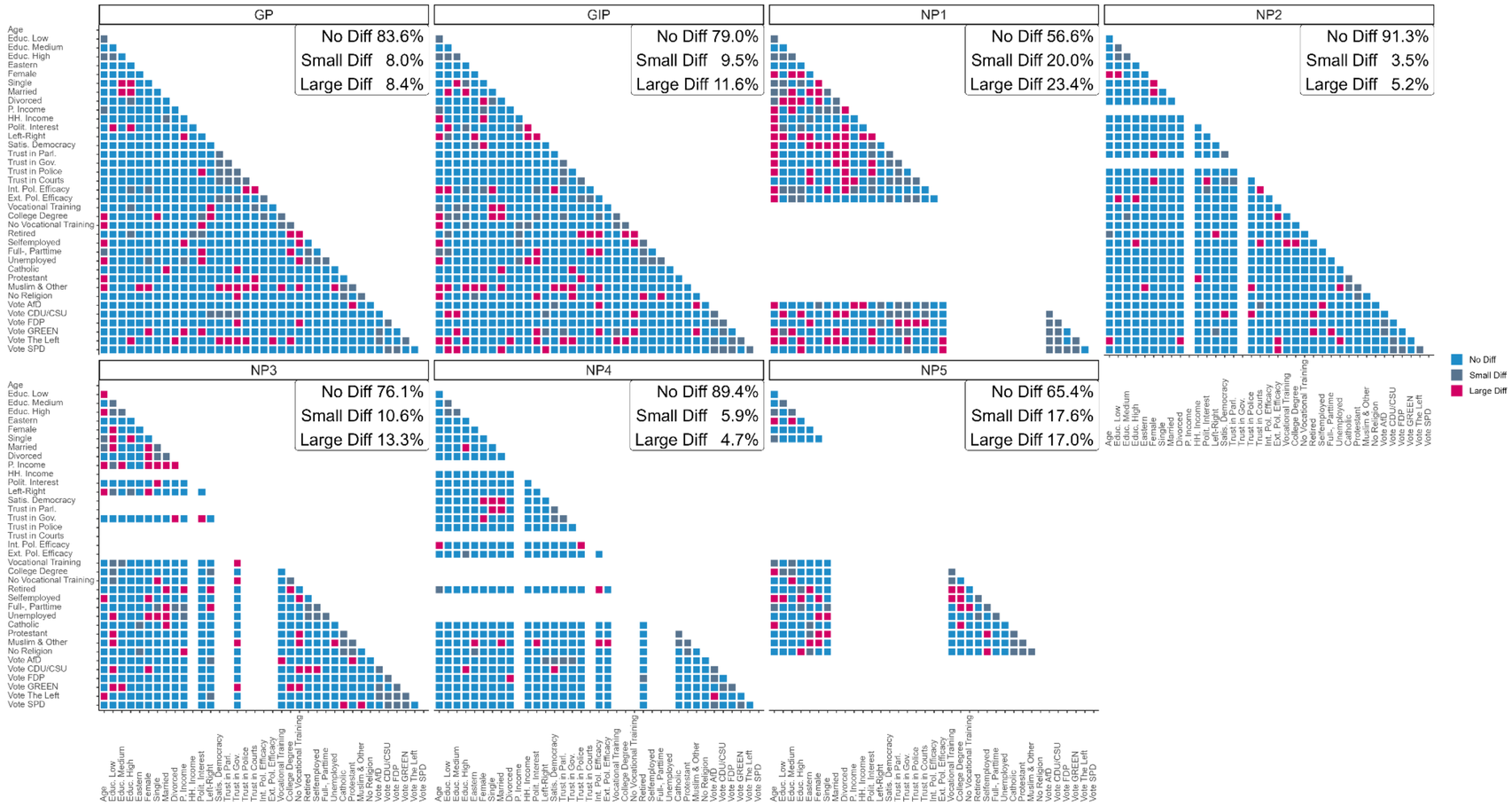
Bias in Univariate Estimates of Probability and Nonprobability Surveys (after Raking)





- Nonprobability survey estimates are, on average, more biased than estimates of probability surveys.
- Variance in RMSE is far higher for nonprobability surveys than for probability surveys.

Bias in Bivariate Estimates of Probability and Nonprobability Surveys (after Raking)



- Many bivariate estimates were unbiased.
- Variance in bias is far higher for nonprobability surveys than for probability surveys.
- Even surveys from the same panel vendors were very different in bias (NP1 vs NP2)

Bias in Piggybacking Recruitement

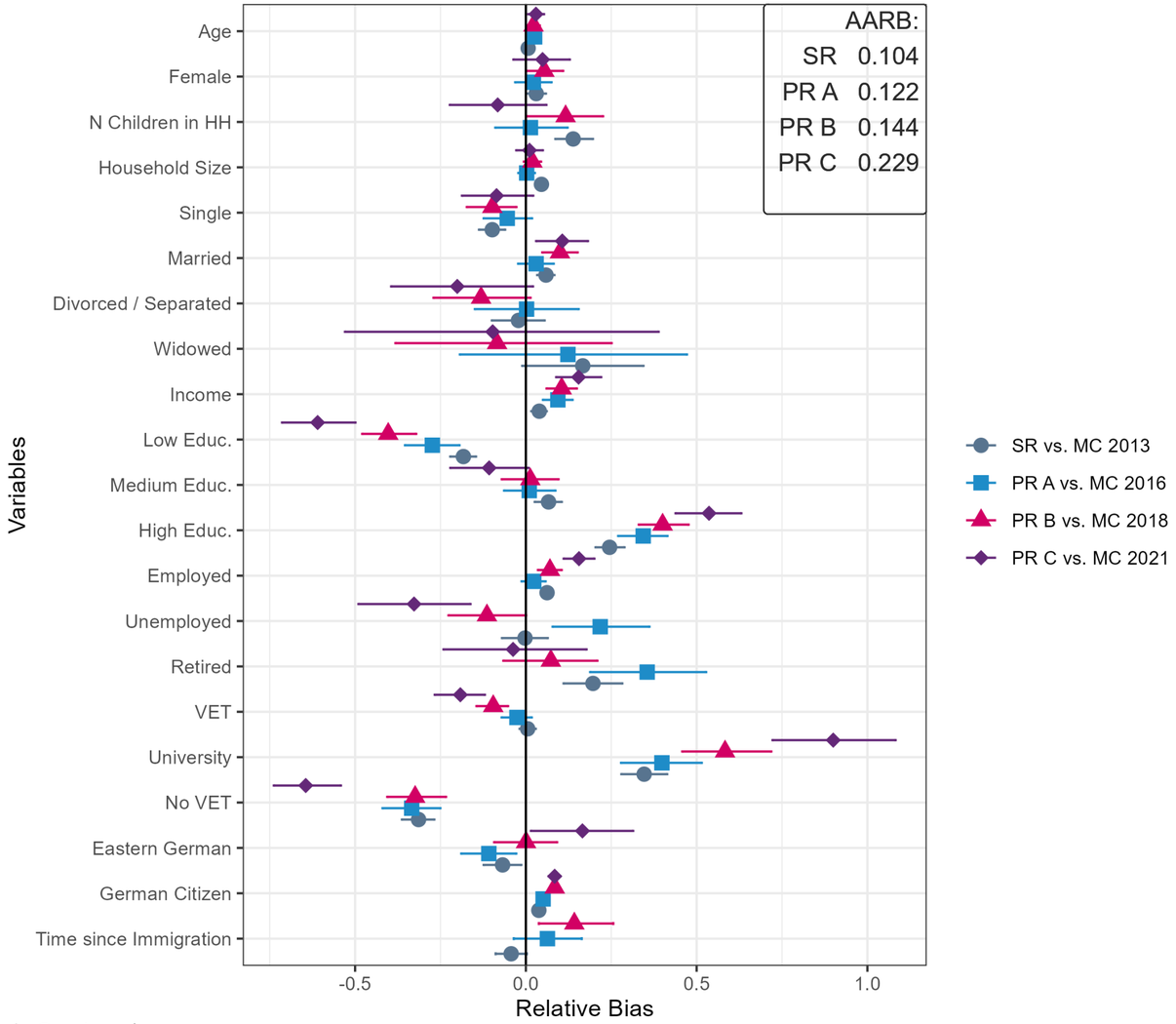
	Separate Recruitment	Piggyback 1	Piggyback 2	Piggyback 3
Year	2013	2016	2018	2022
Piggybacked from:	/	GGSS/ISSP	GGSS/ISSP	ISSP
Sample Size	21,425	10,505	11,334	4,618
N Recruitment Survey	7,599 (35.5%)	/	/	/
N Parent Survey	/	3,490 (33.2%)	3,477 (30.7%)	1,702 (36.9%)
N Agreed to Panel	6,210 (29.0%)	2,124 (20.2%)	2,074 (18.4%)	962 (20,1%)
N Recruited	4,961 (23.2%)	1,710 (16.3%)	1,607 (14,2%)	764 (16.6%)

Note: RR1 in Parenthese

Univariate Bias - Recruitment

SR = Separate Recruitement

PR = Piggyback Recruitement



- Bias is rather similar between SR and PR A and B
- PR C shows higher differences; however, the benchmark had higher nonresponse this year due to the epidemic and technical problems.

GESIS – Leibniz Institute for the Social Sciences

GESIS - Services

GESIS offers individual consulting and other services in a number of areas – including survey design & methodology, data archiving, digital behavioral data & computational social science – and across the research data cycle.

Please visit our website www.gesis.org for more detailed information on available services and terms.

Upcoming Meet the Experts talks:

Meet the Experts - Expert Insights into current Research Topics (gesis.org)

Interested in data quality training, tools and research stays?

Learn more about the KODAQS (Competence Center Data Quality in the Social Sciences) project and join our mailing list:

Thank you for your attention!

Contact: Bjoern.rohr@gesis.org

GitHub

CRAN

Leibniz Association